Delia Georgeta Bekesi Sorana Săveanu

Statistică aplicată în științele sociale



Presa Universitară Clujeană

Delia Georgeta Bekesi Sorana Săveanu

Statistică aplicată în științele sociale

Presa Universitară Clujeană

2021

Referenți științifici

Prof. univ. dr. habil. Sergiu Bălțătescu

Prof. univ. dr. habil. Florica Ștefănescu

ISBN: 978-606-37-1279-1

Universitatea Babeș-Bolyai
Presa Universitară Clujeană
Director: Codruța Săcelean
Str. Hașdeu nr. 51
400371 Cluj-Napoca, România
Tel/Fax: (+40)-264-597.401
E-mail: editura@editura.ubbcluj.ro
http://www.editura.ubbcluj.ro/

Cuprins

PREFAȚĂ

Lucrarea de față reprezintă un punct de pornire în studiul statisticii aplicate în științele sociale și se adresează în primul rând studenților, dar și tuturor celor interesați de îmbunătățirea competențelor în acest domeniu.

Fiecare capitol din carte cuprinde în prima parte o prezentare a aspectelor teoretice, urmate de exemple și exerciții rezolvate. La finalul fiecărui capitol sunt sintetizate principalele concluzii și sunt propuse exerciții și aplicații în scopul verificării cunoștințelor dobândite.

Cartea se dorește a fi un manual practic de statistică adresat tuturor celor care doresc pe de o parte să acumuleze informațiile necesare analizei și prelucrărilor statistice, iar pe de altă parte să aplice și să își testeze cunoștințele. Lucrarea este structurată pe nouă capitole în care sunt detaliate cele două tipuri de statistică: descriptivă și inferențială. Ultimul capitol cuprinde teste grilă propuse pentru fixarea cunoștințelor teoretice și practice, aplicații de sinteză rezolvate și aplicații de sinteză propuse.

În primul capitol este definită statistica, sunt prezentate conceptele de bază ale statisticii sociale și este realizată distincția între conceptele de ”parametru” și "statistică la nivel de eșantion”. Al doilea capitol se concentrează pe clasificarea variabilelor și a nivelurilor de măsurare, gruparea nivelurilor de măsurare și identificarea operațiilor permise pentru fiecare nivel de măsurare. Următorul capitol cuprinde aspecte privind realizarea tabelelor de frecvență și prezentarea grafică a datelor. În continuare sunt prezentați principalii indicatori utilizați în analiza statistică descriptivă. Un capitol este dedicat prezentării indicatorilor tendinței centrale: modul, mediana și media, a proprietăților mediei aritmetice, a pozițiilor relative ale indicatorilor tendinței centrale, precum și a măsurilor localizării. Partea de statistică descriptivă se încheie cu prezentarea celor mai utilizați indicatori ai variației: amplitudinea, abaterea interquartilă, abaterea medie absolută, varianța, abaterea standard și coeficientul de variație. Patru capitole din carte tratează aspecte importante ale statisticii inferențiale: probabilități și distribuția normală, proprietățile distribuției normale, problematica cercetărilor pe eșantioane, intervale de încredere (de confidență), aspecte privind reprezentativitatea, proceduri de eșantionare, teste de semnificație parametrice și non-parametrice și asocierea variabilelor calitative.

Despre autori

Delia Georgeta BEKESI este lector universitar la Universitatea din Oradea, Facultatea de Științe Socio-Umane. Este licențiată în Economie și are masterat în Gestiunea Întreprinderilor Patrimoniale. A obținut titlul de doctor în Sociologie în anul 2011. De-a lungul perioadei de activitate a fost implicată în implementarea unor proiecte de cercetare, precum: Dezvoltare durabilă și calitatea vieții. Studiu de caz în județul Bihor; SocioPlus - Servicii de pregătire, documentare si acces pentru

studenți în programe de licența și masterat în Sociologie și Asistență Socială; SmartDoct – Programe de înaltă calitate pentru doctoranzii și cercetătorii postdoctorat ai Universității din Oradea pentru creșterea relevanței cercetării și inovării în contextul economiei regionale. Principalele domenii de expertiză sunt: managementul, statistica, economia, ocuparea forței de muncă, șomajul. Printre direcțiile recente de cercetare se numără următoarele teme: piața forței de muncă, managementul resurselor umane, microeconomie și salarizarea personalului, temele orizontale ale FSE. Este autor și co-autor a mai multor publicații de specialitate cu teme din domeniile specificate anterior.

Sorana SĂVEANU este lector dr. la Facultatea de Științe Socio-Umane din cadrul Universității din Oradea. Ea este implicată în activități de cercetare din 2003, având un background consistent în realizarea anchetelor sociologice și în analiza statistică a datelor. Ea este specializată în metodologia cercetării sociologice și cercetarea aplicată, a fost membră în numeroase echipe de cercetare pentru diverse proiecte naționale și internaționale și are o vastă experiență în organizarea, planificarea și coordonarea activităților de cercetare sociologică. Începând din 2016, alături de activitatea de cercetare științifică, desfășoară activități didactice având discipline precum Prelucrarea informatizată a datelor sociale, Educație și societate, Managementul sistemelor informaționale, Managementul proiectelor sociale. Rezultatele activităților de cercetare sunt prezentate în peste 20 de articole publicate în reviste de specialitate, 18 capitole și studii în volume colective și peste 35 de lucrări prezentate la conferințe internaționale.

CE ESTE STATISTICA?
CONCEPTE DE BAZĂ ALE STATISTICII SOCIALE

Obiective

Înțelegerea conceptelor: statistica, statistica descriptivă, statistica inferențială

Realizarea distincției între statistica descriptivă și statistica inferențială

Definirea conceptelor de bază ale statisticii sociale

Realizarea distincției între parametru și statistică la nivel de eșantion

Cum definim statistica

Statistica este știința care oferă sens datelor.

Statistica colectează date, le prelucrează, analizează și interpretează. Datele sunt obținute prin diferite proceduri de măsurare, iar prin analiza și interpretarea lor sunt transformate în informații. Informațiile sunt cele care aduc un plus de cunoaștere celui care le utilizează.

Statistica ne oferă un plus de cunoaștere asupra realității sociale. Informațiile se bazează pe rezultatele obținute prin aplicarea instrumentelor de măsurare. Măsurarea presupune observarea anumitor caracteristici ale fenomenelor și proceselor analizate.

Statistica se referă pe de o parte la modul în care sunt culese datele, la modul în care sunt descrise acestea și, pornind de la rezultatele obținute în urma analizei acestora, contribuie la explicarea proceselor și fenomenelor analizate.

Primul pas în analiza statistică este descrierea datelor măsurate. Apoi, pe baza rezultatelor obținute, extragem niște concluzii care se răsfrâng asupra altor date care nu au fost măsurate inițial. Aceasta este diferența dintre statistica descriptivă și statistica inferențială.

Statistica descriptivă cuprinde ansamblul procedurilor statistice prin care se descrie din punct de vedere statistic o populație sau un eșantion. Statisticile descriptive oferă o imagine sintetică asupra datelor.

Statistica inferențială cuprinde ansamblul procedurilor statistice prin care se fac generalizări privind caracteristicile populației pornind de la un eșantion. Statisticile inferențiale permit formularea unor concluzii referitoare la întreaga populație pornind de la realizarea unor măsurători doar asupra unei părți din populații. Acesta este rolul cel mai important al statisticii, deoarece de cele mai multe ori nu este posibil (din cauza resurselor disponibile – resurse financiare, umane și de timp) să

culegem date de la toată populația. Prin diferite proceduri statistice, datele culese de la o parte din populație produc rezultate cu referință la ansamblu.

 Exemplu

Să presupunem că la nivel național se dorește introducerea uniformelor în școli pentru elevii din clasele 9-12. Dar, înainte de a introduce această prevedere, Ministerul educației dorește să afle în ce măsură elevii sunt de acord cu aceasta. Este dificil să aflăm răspunsul fiecărui elev în parte, ca atare, vom selecta aleator doar o parte din elevi, să zicem 2000 de elevi. Să presupunem că la întrebarea *Ești de acord ca purtarea uniformei în școală să fie obligatorie în clasele 9-12?*, 72% dintre elevi au răspuns *Nu*. Pe baza inferențelor statistice, vom putea să concluzionăm că majoritatea elevilor nu doresc să fie introdusă purtarea uniformelor în școală. Totuși, pentru a formula astfel de concluzii este foarte important cum am selectat indivizii și să ne asigurăm că am măsurat ceea ce ne-am propus să măsurăm.

Concepte de bază ale statisticii sociale

Populația statistică reprezintă mulțimea indivizilor statistici care prezintă interes pentru studiu (suma unităților de analiză). Cercetările care cuprind toată populația statistică se numesc *cercetări exhaustive*.

Unitatea statistică (unitatea de analiză, individul statistic, observația statistică) reprezintă fiecare element component al populației statistice.

Eșantionul este un subset sau submulțime a populației statistice. Cercetările realizate pe eșantioane se numesc cercetări selective.

Variabila statistică este orice caracteristică a membrilor unei populații (sau a unui eșantion), care variază în respectiva populație (sau eșantion). Fiecare variabilă poate lua anumite valori.

 Exemple de variabile

Într-o populație statistică poate varia variabila Starea civilă.
Valorile pe care le poate avea variabila sunt: căsătorit, necăsătorit, văduv, divorțat, separat (despărțit nelegal)
O altă variabilă poate să fie religia. Valorile care variază sunt: Ortodoxă, Greco-catolică, Romano-catolică, Protestantă, Iudaică, Musulmană, Neo Protestanta, sau altă religie.
Într-o populație poate varia numărul de copii din gospodărie. Valorile variabilei pot fi: 0, 1, 2, 3 etc..
Alte variabile pot fi: înălțimea, venitul/lună, statusul ocupațional, vârsta, opțiunea la vot, nivelul de instrucție etc..

Orice cercetare sociologică, indiferent de subiect, va avea o populație statistică, o unitate statistică și variabile ale populației. În plus, de cele mai multe ori, cercetările sociologice sunt realizate pe eșantioane (datorită consumului redus de resurse).

Distincția între parametru și statistică la nivel de eșantion

Parametrul este valoarea unei variabile la nivelul populației statistice (exemplu: vârsta medie a studenților calculată la nivelul populației statistice).

Statistica la nivel de eșantion se referă la valoarea unei variabile la nivel de eșantion. (exemplu: vârsta medie a studenților calculată pe un eșantion).

Statistica inferențială estimează valoarea parametrilor.



Exemple

Pentru o cercetare cu privire la consumul de droguri în rândul studenților Universității din Oradea vom avea:
Populația statistică = toți studenții Universității din Oradea
Eșantionul = o parte selectată din mulțimea studenților Universității din Oradea
Unitatea statistică = studentul
Variabila statistică = frecvența consumului de droguri

O cercetare cu privire la performanțele școlare ale elevilor bihoreni din ciclul învățământului liceal va cuprinde:
Populația statistică = totalitatea elevilor de liceu din Bihor
Eșantionul = o submulțime a elevilor de liceu din Bihor
Unitatea statistică = elevul de liceu
Variabila statistică = rezultate școlare (media anului școlar)
Vom înregistra toate datele la nivelul eșantionului; fiecare medie din anul școlar obținută de fiecare elev selectat în eșantion
Un set de date = mediile din anul școlar ale elevilor liceeni
Parametru = valoarea medie a mediei estimată la nivelul populației elevilor de liceu din Bihor pornind de la valoarea medie a mediei înregistrată la nivelul eșantionului.

Concluzii

Statistica presupune măsurare și analiza probabilităților.
Statistica descriptivă are ca scop sumarizarea și prezentarea datelor.
Statistica inferențială are ca scop efectuarea unor generalizări despre o populație statistică
Conceptele de bază ale statisticii sunt: populație statistică, individ statistic, eșantion, variabilă, parametru, statistică la nivel de eșantion.

✍ *Exerciții și aplicații*

Identificați populația statistică, unitatea de analiză și variabila, dacă obiectul studiului îl reprezintă:

1. Distribuția opțiunilor electorale
2. Performanțele școlare ale elevilor
3. Gradul de poluare a localităților urbane
4. Antreprenorialul în rândul femeilor

TIPURI DE VARIABILE ȘI NIVELURI DE MĂSURARE

Obiective

Prezentarea nivelurilor de măsurare. Gruparea nivelurilor de măsurare. Tipuri de variabile

Identificarea operațiilor statistice permise pentru fiecare nivel de măsurare

Niveluri de măsurare

Sunt patru niveluri de măsurare: nominal, ordinal, de interval și de raport. În funcție de nivelul de măsurare identificăm tipul variabilelor.

Nivelul nominal presupune clasificarea în categorii între care nu există o relație de ordine. Categoriile variabilei trebuie să fie *mutual exclusive* (categorii distincte, indivizii statistici nu pot fi clasificați simultan în mai multe categorii)și *exhaustive* (să existe toate valorile posibile pe care le poate lua o variabilă; în cazul variabilelor care au un număr foarte mare de categorii se pune și varianta ”alta”). Cu alte cuvinte, fiecare individ statistic trebuie să se regăsească obligatoriu într-o categorie, dar doar în una singură.

 Exemple

Starea civilă:
1. Necăsătorit 2. Căsătorit, coabitare 3. Divorțat 4. Văduv 5. Separat (despărțit nelegal)

În chestionare de obicei folosim coduri pentru valorile variabilei (pentru variantele de răspuns). Aceste cifre sunt doar simboluri. În exemplul de mai sus nu putem ordona valorile. 2 nu este mai mare decât 1 deoarece Căsătorit nu este mai mare decât Necăsătorit, la fel cum nici 3 nu este mai mare decât 1, deoarece Divorțat nu este mai mare decât Necăsătorit. Este absurd să punem o ordine între valorile unei variabile la nivel nominal.

Gândindu-vă la ultimele alegeri parlamentare, ce metodă de campanie electorală v-a plăcut cel mai mult ?
1. Afișul
2. Bannere/panourile
3. Articole din presă și emisiuni TV
4. Acțiunile de stradă-materialele primite
5. Discuțiile, întâlnirile cu candidații
6. Mesajele, discuțiile pe internet
7. Materialele primite prin poștă
8. Altă metodă

 Exemple

La fel ca în exemplul de mai sus, nu există o ordine între valorile variabilei. Variantele de răspuns sunt codificate doar pentru a fi mai ușor să le gestionăm în baza de date.

Locuința în care stați în prezent este …:

1. Proprietatea dvs. / partenerului(erei) dvs.
2. Proprietatea părinților (rudelor)
3. Închiriată de la o persoană / firmă
4. Închiriată de la stat
5. Locuință socială
6. Locuință de serviciu

Pe când aveați 14 ani, locuiați împreună cu părinții?

1 – da, cu amândoi părinții
2 – da, doar cu tata
3 – da, doar cu mama
4 – nu, nu am locuit cu părinții

La *nivelul ordinal* categoriile variabilei pot fi ordonate ierarhic. Astfel, între categoria A și B ale variabilei vom avea fie A<B, fie A>B, și A≠B. Totodată este respectată proprietatea de tranzitivitate: dacă A<B și B<C, atunci A<C.

 Exemple

Avem următoarea întrebare într-un chestionar: Vă rugăm să ne spuneți cât de importantă este familia în viața dvs.?

Variantele de răspuns sunt:

4. Foarte importantă 3. Destul de importantă 2. Puțin importantă 1. Deloc importantă

Nivelul de măsurare pentru această variabilă este cel ordinal deoarece între valorile variabilei (variantele de răspuns la întrebare) există o ordine. Cifrele folosite pentru variantele de răspuns sunt doar coduri, dar, față de variabilele măsurate la nivel nominal, acum le putem ordona. 1<2<3<4. Există o ordine între cele 4 variante de răspuns. De exemplu, știm că pentru subiecții care au răspuns "Puțin importantă" familia este mai puțin importantă decât pentru cei care au răspuns "Destul de importantă".

Alte exemple de variabile măsurată la nivel ordinal:

De obicei cât de des petreceți timpul în afara orelor de program, cu colegii de serviciu sau de profesie?

4. cel puțin o dată pe săptămână 3. o dată sau de două ori pe lună
2. de câteva ori pe an 1.deloc

Cât de mulțumit sunteți de activitatea Guvernului în domeniul protecției mediului?
4. Foarte mulțumit 3. Mulțumit 2. Nemulțumit 1. Foarte nemulțumit

Cum apreciați situația zonei în care locuiți în ceea ce privește starea drumurilor?
5. Foarte bună 4. Bună 3. Satisfăcătoare 2. Proastă 1. Foarte proastă
În ce măsură credeți că oameni ca dvs. pot influența hotărârile importante care se iau pentru localitatea dvs.?
4. În foarte mare măsură 3. În mare măsură 2. În mică măsură

1. În foarte mică măsură/ Deloc
În afară de nunți, înmormântări și botezuri, cât de des ați mers în ultimul timp la biserică?
1 – De mai multe ori pe săptămână
2 – O dată pe săptămână
3 – O dată pe lună
4 – Doar la cele mai importante sărbători religioase
5 – O dată pe an
6 – O dată la câțiva ani
7 – Niciodată sau aproape niciodată

Variabilele măsurate la nivel nominal și ordinal sunt incluse în categoria variabilelor calitative**.** Diferența dintre categoriile (valorile) variabilei este una calitativă.
La *nivelul de interval* valorile variabilei pot fi ordonate, există intervale egale între valori, dar nu există o origine sau un zero absolut. "Zero" în cazul acestei variabile este unul convențional (nu există în realitate, ci este stabilit de către cercetător).

 Exemple

La un examen, putem să evaluăm studenții folosind o scală cu note de la 2 la 10 (să presupunem că le oferim 2 puncte din oficiu ☺).
În acest caz, profesorul stabilește scala de evaluare. El va ști că studentul care a luat nota 3 este mai slab pregătit decât studentul care a luat nota 9 (există o ordine între valori). Dar punctul de unde începe notarea (zero absolut) este stabilit de către profesorul evaluator. La fel de bine poate să folosească o scală de notare de la 12 la 20. Important este să știe că studentul care a luat nota 12 este mai slab pregătit decât cel care a luat nota 20.
Scala de măsurare a temperaturii – grade Celsius sau Fahrenheit
Scala de măsurare a Nivelului de inteligență IQ.

Față de nivelul de interval, *nivelul de raport* presupune existența unui zero absolut (există în mod natural). În plus, acest nivel include toate caracteristicile nivelurilor

anterioare. Acest nivel include toate variabilele în cazul cărora cifrele reprezintă într-adevăr numere.

 Exemple

Numărul de copii din gospodărie: 0, 1, 2, 3, 4 etc.. 0 nu este stabilit de către cercetător. El există în realitate și semnifică "în gospodărie nu sunt copii".
Venitul din gospodărie, vârsta, numărul de kilometri de autostradă realizați în anul 2018, numărul de tichete de parcare vândute într-o lună, numărul de studenți participanți la conferință, numărul de beneficiari ai serviciilor sociale, numărul de ore de meditație la matematică.

Cele două niveluri de măsurare (de interval și de raport) formează categoria variabilelor numerice. Valorile variabilelor sunt numere, iar diferența dintre ele este una numerică.

În plus, aceste variabile numerice pot fi clasificate în alte două categorii:

Variabile discrete: care iau valori întregi (nu pot lua decât anumite valori)
Variabile continue: care iau valori reale, iar între două valori succesive ale unei variabile pot exista o infinitate de valori

 Exemple

Variabile discrete: numărul de copii din gospodărie. Nu pot în gospodărie să fie 2.3 copii, sunt fie 2, fie 3 copii.
Alte exemple: numărul de bancnote din portmoneu, numărul de pachete de făină de pe raft.
Variabile continue: vârsta (numărul de persoane înregistrat pe categorii de vârstă).
Numărul salariaților din industrie în județul Bihor: Număr total 2078.

	din care: pe clase de mărime, după numărul mediu al salariaților			
	0-9	10-49	50-249	250 și peste
Industrie	1423	454	164	37

Iată în sinteză informația prezentată în această secțiune: clasificăm variabilele în funcție de posibilitatea de ordonare, de existența unității de măsură și de existența unui zero absolut. Aceste trei criterii ne indică operațiile matematice care sunt permise și care sunt interzise pentru fiecare nivel de măsurare.

Nivelul de măsurare	**Posibilitate de ordonare**	**Unitate de măsură**	**Zero absolut**	**Operații matematice premise**
NOMINAL	NU	NU	NU	nici una!
ORDINAL	**DA**	NU	NU	Ordonare
DE INTERVAL	**DA**	**DA**	NU	Ordonare, Scădere, Adunare
DE RAPORT	**DA**	**DA**	**DA**	Ordonare, Scădere Adunare, Înmulțire împărțire

Variabile dihotomice (dummy)

Un caz particular de variabile calitative este cel al variabilelor dihotomice. Aceste variabile presupun clasificarea indivizilor statistici în două categorii. Variabila are doar două valori, iar fiecare individ ia una dintre aceste valori. De obicei aceste variabile folosesc codurile 0 și 1.

 Exemple

Genul subiecților:
1. masculin   2. feminin
Fiecare subiect va putea fi inclus într-una din cele două categorii, fie a bărbaților, fie a femeilor.
Sunteți de acord cu introducerea unei taxe de campare în zona Glăvoi (Parcul Natural Apuseni)? 1. Da   0. Nu
Sunteți cetățean român? 1. Da   0. Nu
În ultimele 12 luni, dvs. ați participat la vreo întrunire publică legată de problemele zonei? 1. Da   0. Nu

Gruparea nivelurilor de măsurare

Gruparea variabilelor în cele două categorii menționate anterior (calitative și numerice) fac distincția dintre *statistica non-parametrică* și *statistica parametrică*. Statistica non-parametrică se aplică variabilelor calitative, iar cea parametrică variabilelor numerice. Se poate aplica statistica non-parametrică și variabilelor cantitative, nu și invers.

Statistica non-parametrică	*Statistica parametrică*
Nivel nominal și ordinal	Nivel de interval și de raport
Variabile calitative (categoriale, non parametrice)	Variabile numerice (cantitative, numerice, parametrice)

Cunoscând tipul de variabilă pe care îl avem, știm ce fel de analize pot fi aplicate setului de date de care dispunem. Nivelurile de măsurare sunt importante pentru clasificarea și etichetarea variabilelor într-un studiu, dar și pentru elaborarea întrebărilor dintr-un studiu.

Operații statistice permise pentru fiecare nivel de măsurare

Nivelul nominal permite calcularea frecvențelor (absolute, relative și cumulate), valoarea modală (modul), verificarea statistică prin chi-pătrat de concordanță, calculul coeficienților de contingență (asociere)

Nivelul ordinal. Pe lângă calculele anterioare, pentru variabile ordinale se mai pot determina: mediana și indicatorii înrudiți (centile, decile, quartile, abatere interquartilă) precum și coeficienți de corelație a rangurilor.

Nivelul de interval. Acest nivel oferă în plus față de nivelul anterior și calculul mediei, precum și calcule legate de ea: abaterea medie, coeficient de corelație și regresie, testul t, analiza de varianță.

Nivelul de raport. La acest nivel sunt permise toate operațiile statistice, inclusiv calculul mediei geometrice și al coeficientului de variație.

Indicatori statistici ce pot fi calculați în funcție de nivelul de măsurare:

Nivel de măsurare (Tip variabilă)	Indicatori ai tendinței centrale		
	Modul	Mediana	Media
Nominal	da	-	-
Ordinal	da	da	-
De interval	da	da	da
De raport	da	da	da

Nivel de măsurare (Tip variabilă)	Indicatori ai dispersiei						
	Quartile	Abaterea interquartilă	Amplitudine	Abatere medie	Abaterea medie absolută	Varianța	Abatere standard
Nominal	-	-	-	-	-	-	-
Ordinal	da	da	-	-	-	-	-
De interval	da	da	da	da	da	da	da
De raport	da	da	da	da	da	da	da

Nivel de măsurare (Tip variabilă)	Indicatori ai formei distribuției	
	Oblicitate	Boltire
Nominal	-	-
Ordinal	-	-
De interval	da	da
De raport	da	da

Concluzii

Putem identifica patru niveluri de măsurare: nominal, ordinal, de interval, de raport.

Cele 4 niveluri de măsurare generează 2 mari tipuri de variabile: variabile calitative (non-parametrice, categoriale) și variabile cantitative (numerice, parametrice).

Variabilele numerice pot fi discrete sau continue (intervalul dintre două valori succesive poate fi divizat la infinit).

Variabilele dihotomice sunt cele care au două valori.

Tipul variabilelor indică operațiile statistice permise și cele nepermise.

✍ *Exerciții și aplicații*

Exercițiul 1.

Dați 3 exemple (altele decât cele din acest capitol) de variabile măsurate la nivel nominal, respectiv ordinal.

Dați 2 exemple (altele decât cele din acest capitol) de variabile măsurate la nivel de interval, respectiv de raport.

Dați 3 exemple (altele decât cele din acest capitol) de variabile dihotomice.

Exercițiul 2. Identificați nivelul de măsurare pentru următoarele variabile:

A. Activitatea desfășurată la locul de muncă are legătură cu studiile absolvite?

4. Întotdeauna 3. În majoritatea situațiilor 2. Rareori 1. Niciodată

B. Ce proporție din cunoștințele acumulate în timpul facultății folosiți la locul de muncă? ……. %

C. Care este modalitatea ta principală de conectare la cursurile online?

1. Laptop 2. Telefon mobil 3. PC (Desktop) 4. Tabletă 5. Altă modalitate

D. În lista următoare găsiți mai multe afirmații despre materialele didactice. Vă rugăm să apreciați pe o scală de la 1 la 10 cât de adevărate sunt afirmațiile:

	Deloc adevărat									Perfect adevărat
Înțeleg informațiile prezentate în suportul de curs	1	2	3	4	5	6	7	8	9	10
Ni s-au prezentat la începutul cursurilor obiectivele, tematica, bibliografia și cerințele	1	2	3	4	5	6	7	8	9	10
Temele au fost prezentate într-o succesiune logică	1	2	3	4	5	6	7	8	9	10

E. În medie, în ultimele 3 luni, cât de frecvent ați desfășurat activitate sportivă de cel puțin 30 de minute?
1. niciodată
2. 1 data/luna sau mai rar
3. de mai multe ori/luna
4. de 1-2 ori pe săptămână
5. de 3 sau mai multe ori/săptămână
6. zilnic

DISTRIBUȚII DE FRECVENȚE. REPREZENTĂRI GRAFICE

Obiective

Înțelegerea elementelor de bază ale unei distribuții statistice

Calcularea frecvențelor relative și a frecvențelor cumulate

Cunoașterea celor mai utilizate grafice atașate distribuțiilor de frecvență

Realizarea tabelelor de frecvență

O distribuție statistică (de frecvențe) reprezintă un set de date sistematizat și organizat care conține valorile (categoriile) variabilei și frecvențele absolute. De regulă, o distribuție de frecvențe este prezentată sub forma unui tabel.

Frecvența absolută (f_i) reprezintă numărul de indivizi statistici (observații, unități de analiză) înregistrat pe fiecare categorie a variabilei.

$\sum f_i = N$ N- numărul total de indivizi statistici

Pornind de la frecvențele absolute putem calcula frecvențele relative și frecvențele cumulate.

Frecvența relativă (fr_i) reprezintă ponderea (proporția; procentul) indivizilor statistici din fiecare categorie raportată la numărul total al indivizilor statistici și se determinată cu ajutorul formulei de mai jos:

$$fr_i = \frac{f_i}{\sum f_i} * 100$$

unde fi este frecvența absolute
și $\sum f_i = N$ reprezintă numărul total

Frecvența cumulată se determină prin însumarea frecvenței respectivei categorii, și frecvențele categoriilor inferioare.

Exemple

Exercițiul 1. Am obținut de la 23 de studenți următoarele răspunsuri la întrebarea: *De obicei, în cursul săptămânii, unde iei masa de prânz?*

"la cantină"	*"în oraș"*	*"în oraș"*
"acasă"	*"în oraș"*	*"în oraș"*
"acasă"	*"la cantină"*	*"acasă"*
"în altă locație"	*"în altă locație"*	*"acasă"*
"în oraș"	*"la cantină"*	*"la cantină"*
"acasă"	*"acasă"*	*"în oraș"*
"acasă"	*"acasă"*	*"acasă"*
"în oraș"	*"la cantină"*	

Pentru a realiza tabelul de frecvență, în primul rând trebuie să identificăm valorile variabilei. Avem o variabilă măsurată la nivel nominal, iar valorile (categoriile de răspuns) sunt: "la cantină" , "în oraș", "acasă", "în altă locație".

Apoi, pentru fiecare valoare (categorie) în parte trebuie să aflăm frecvența absolută (numărul de răspunsuri pentru fiecare categorie). Dacă vom lua valoarea (categoria) "la cantină", vom vedea că din cei 23 de studenți care au răspuns la întrebarea noastră, 5 au ales varianta "la cantină". 5 este frecvența absolută a valorii "la cantină". În același mod vom afla frecvențele absolute pentru toate categoriile variabilei.

Tabelul de frecvență se prezintă astfel:

Valoarea variabilei x	**Frecvențe absolute f_i**	**Frecvențe relative (%) fr_i**	**Frecvențe cumulate**
la cantină	5	21.74	21.74
în oraș	7	30.43	52.17
acasă	9	39.13	91.30
în altă locație	2	8.70	100.00
	N=23	100	

Numărul total de cazuri/ numărul total de indivizi statistici = suma frecvențelor absolute

21.74 este frecvența relativă pentru categoria "la cantină", care are frecvența absolută 5. Se calculează după formula:

$\frac{5}{23}$ *100=21.74

Raționamentul este următorul: din 23 de studenți, care reprezintă 100%, valoarea 5 (5 studenți au răspuns "la cantină"), ce procent reprezintă?

Vom interpreta astfel: "Din totalul studenților (N=23), 21.74% declară că își iau prânzul la cantină."

	Frecvențe relative (%) fr_i	Frecvențe cumulate
la cantină	21.74	21.74
în oraș	30.43	52.17
acasă	39.13	91.30
în altă locație	8.70	100.00
	100%	

Pentru a calcula frecvențele cumulate vom porni de la frecvența primei categorii, apoi vom aduna frecvențele categoriilor următoare.
52.17 = 21.74 + 30.43
91.30 = 52.17 + 39.13
100 = 31.30 + 8.70

Exercițiul 2. Am adresat următoarea întrebare unui grup de 108 indivizi: *În medie, în ultima lună, de câte ori ați mers la cinematograf cu prietenii?*
Tabelul de frecvență se prezintă astfel:

Valoarea variabilei x	Frecvențe absolute f_i	Frecvențe relative (%) fr_i	Frecvențe cumulate
1	31	28.70	28.70
3	24	22.22	50.93
4	18	16.67	67.59
5	13	12.04	79.63
6	11	10.19	89.81
7	7	6.48	96.30
9	4	3.70	100.00
	N=108	100%	

Vom interpreta astfel: 31 de indivizi au declarat că au fost 1 dată la cinematograf cu prietenii în ultima lună. Cei 31 de indivizi reprezintă 28.70% din totalul subiecților care au răspuns la întrebare.

$$28.70 = \frac{31}{108} * 100$$

24 de indivizi au declarat că au fost de 3 ori la cinematograf cu prietenii în ultima lună. Cei 24 de indivizi reprezintă 22.22% din totalul subiecților.

$$22.22 = \frac{24}{108} * 100$$

Pentru a calcula frecvențele cumulate vom frecvența relativă a categoriei cu frecvența relativă a categoriei următoare: 50.93 = 27.70+22.22; 67.59 = 50.93+16.67; 79.63=67.59+12.04 etc.. Vom interpreta astfel: 79.63% dintre indivizi merg de până la 5 ori la cinematograf, și anume, de 1, de 3, de 4 și de 5 ori.

Prezentarea grafică a datelor

Reprezentarea grafică a unui tabel de frecvență se face de obicei cu ajutorul diagramelor (bară sau coloană) și a histogramei.

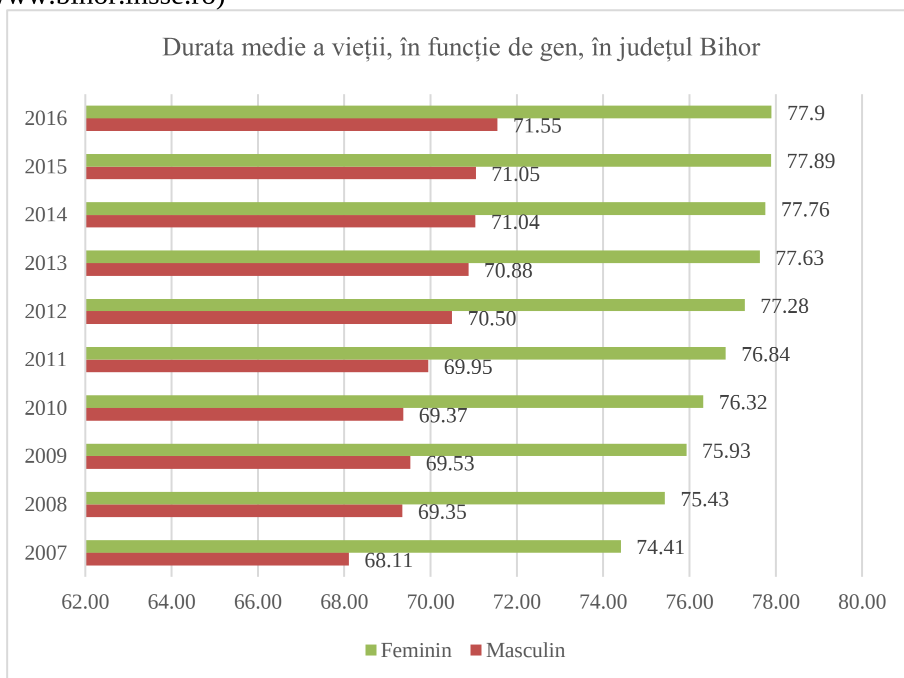
Diagramele sunt utilizate în cazul variabilelor calitative (nominale sau ordinale). În funcție de numărul de categorii ale variabilei, vom alege între o reprezentare grafică tip bare, coloane sau tip plăcintă ("Pie Chart"). Nu este recomandat să folosim un grafic tip plăcintă dacă variabila are multe categorii.

Histograma este utilizată în cazul informațiilor de tip cantitativ. Ea este folosită pentru a arăta care este forma distribuției unei variabile.

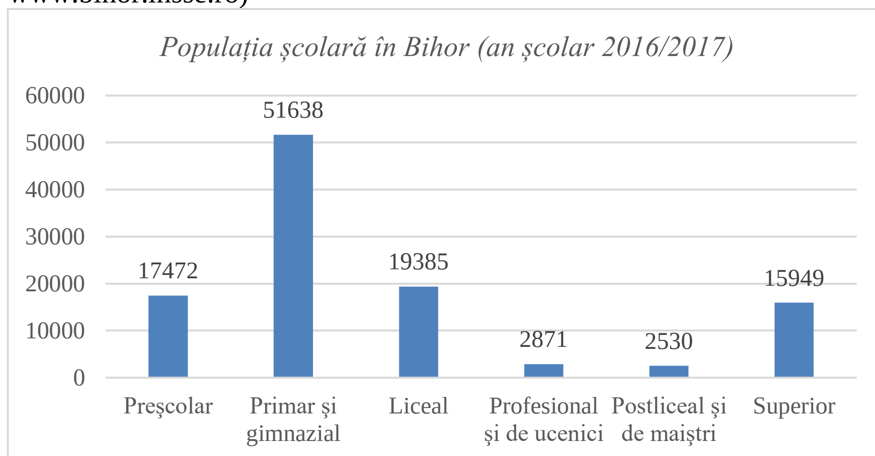
În cazul variabilelor cantitative, în special dacă dorim să reprezentăm evoluția în timp a variabilei, putem să folosim o reprezentare grafică tip linie.

 Exemple

Diagramă tip bare (sursa datelor: Direcția Județeană de Statistică BIHOR, www.bihor.insse.ro)



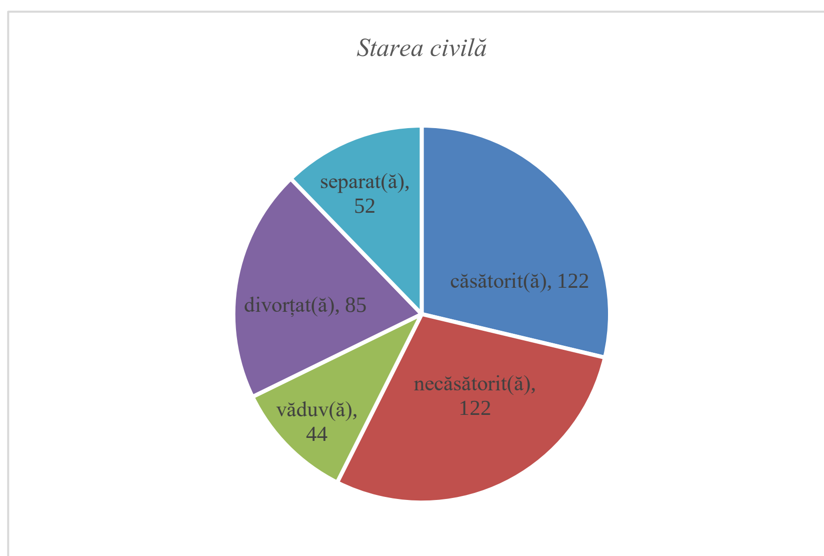
Diagramă tip coloane (sursa datelor: Direcția Județeană de Statistică BIHOR, www.bihor.insse.ro)



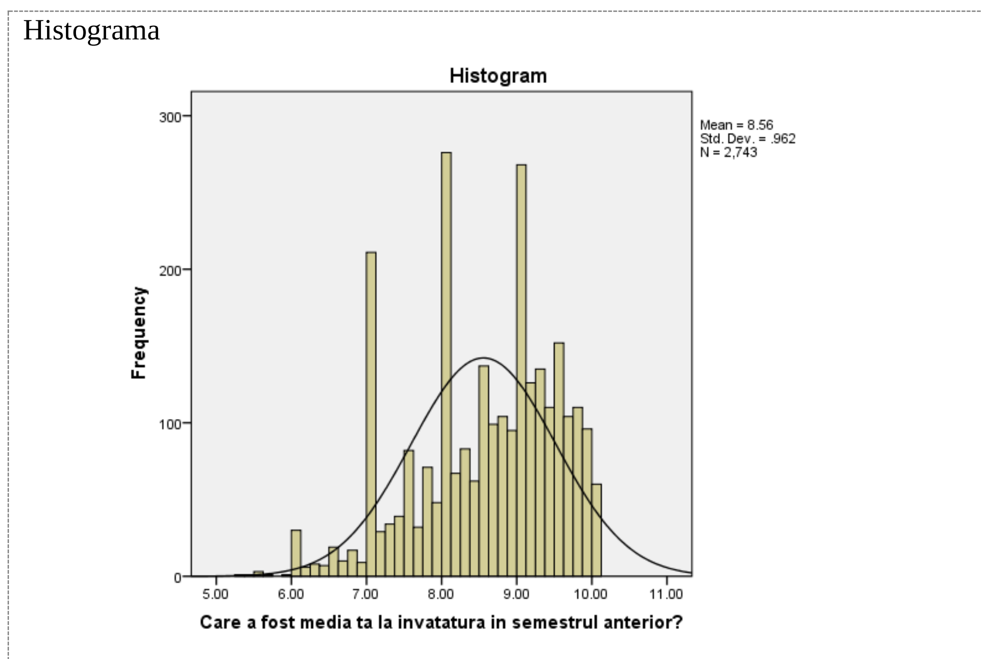


Exemple

Diagramă tip plăcintă (date fictive)



Histograma



Concluzii

O distribuție statistică (de frecvențe) reprezintă un set de date sistematizat și organizat care conține valorile (categoriile) variabilei și frecvențele absolute. Alte informații care mai pot fi incluse sunt frecvențele relative, frecvențele cumulate absolute și relative

Distribuțiile de frecvențe pot fi reprezentate grafic, cele mai utilizate grafice fiind histograma (pentru variabile cantitative) și diagrama bară (pentru variabile calitative)

✍ *Exerciții și aplicații*

Exercițiul 1. În cadrul unei cercetări asupra unui eșantion de șomeri, am obținut următorul tabel de frecvență privind distribuția în funcție de numărul de copii din gospodăriile șomerilor:

Număr de copii din gospodărie	Număr de gospodării (frecvențe absolute)
1	220
2	158
3	30
4	10
5	2
Total	420

Determinați:

Frecvențele relative

Frecvențele relative cumulate

Care este ponderea (proporția) gospodăriilor care au 5 copii?

Care este ponderea (proporția) gospodăriilor care au 4 copii sau mai puțin?

Exercițiul 2. În cadrul unui sondaj am obținut următoarea distribuție în ceea ce privește numărul de elevi înscriși pe clase:

Clasa	Frecvența
cl. a 9-a	32
cl. a 10-a	29
cl. a 11-a	32
cl. a 12-a	31

Calculați frecvențele relative și frecvențele cumulate.

Exercițiul 3. În cadrul unui sondaj am obținut următoarea distribuție în ceea ce privește variabila "starea civilă":

Starea civilă	Frecvența absolută
Necăsătorit	40
Căsătorit	55
Văduv	5
Divorțat	10

Calculați frecvențele relative; interpretați rezultatele obținute

Exercițiul 4. Calculați frecvențele relative și cumulate pentru următoarea distribuție:

În medie, în ultimele 3 luni, cât de frecvent ați desfășurat activitate sportivă de cel puțin 30 de minute?	Frecvența
niciodată	144
1 data/luna sau mai rar	51
de mai multe ori/luna	57
de 1-2 ori pe săptămână	50
de 3 sau mai multe ori/săptămână	41
zilnic	44

INDICATORI AI TENDINȚEI CENTRALE

Obiective

Descrierea indicatorilor tendinței centrale

Prezentarea proprietăților mediei aritmetice

Calculul indicatorilor tendinței centrale

Utilizarea corectă a indicatorilor tendinței centrale în funcție de nivelul de măsurare

Cunoașterea și înțelegerea măsurilor localizării

Modul, mediana și media

MODUL (valoarea modală) este acea valoare a variabilei care apare cel mai des într-un eșantion sau într-o populație. Modul este acea categorie a variabilei care are frecvența cea mai mare.

Dacă reprezentăm grafic distribuția unei variabile, vom identifica modul ca fiind valoarea căreia îi corespunde vârful distribuției.

Când datele sunt grupate pe intervale vom determina intervalul modal (intervalul cu frecvența cea mai mare).

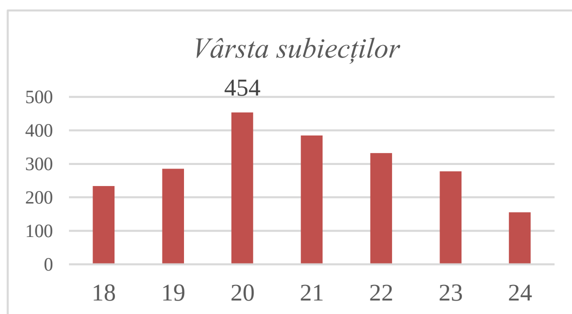
În cadrul cercetărilor pot fi identificate și distribuții bimodale, atunci când 2 valori ale variabilei apar cu o aceeași cea mai mare frecvență.



Exemple

Pentru următorul tabel de frecvență, modul este “20 de ani” deoarece subiecții cu vârsta de 20 de ani au frecvența cea mai mare (454). Ei sunt cei mai numeroși din întregul eșantion.

Vârsta	Nr. de subiecți
18	234
19	285
20	454
21	385
22	332
23	278
24	155



În exemplul de mai jos este prezentată o distribuție bimodală: valorile 2 și 3 înregistrează aceeași frecvență, 42 fiind cifra cea mai mare.

Număr de persoane în gospodărie	Frecvența absolută
1	36
2	42
3	42
4	31
5	24
6	12
7	9

MEDIANA este acea valoare a variabilei care împarte seria ordonată de date în 2 părți egale, astfel încât 50% din observații se vor găsi deasupra valorii mediane, iar 50% se vor găsi dedesubtul ei.



Mediana

Dacă vom calcula frecvențele cumulate, mediana este valoarea corespunzătoare primei frecvenței cumulate care depășește jumătate din observații.

În determinarea medianei trebuie să ținem cont dacă avem un număr par sau impar de observații.

 Exemple

Am obținut următoarele răspunsuri la întrebarea: *În prezent câte persoane locuiesc în gospodăria dvs.?*

1 2 2 3 (3) 3 4 5 5

Numărul este impar; sunt 9 persoane care au răspuns la întrebarea noastră.

M_e=3 deoarece se află la mijloc; 50% dintre observații sunt în dreapta, iar 50% în stânga.

Dacă avem un număr de observații impar, vom calcula media valorilor din mijlocul seriei. Să presupunem că am obținut următoarele răspunsuri la întrebarea: *În prezent câte persoane locuiesc în gospodăria dvs.?*

1 1 2 (2 3) 3 4 5

M_e=(2+3)/2=2.5

Este foarte ușor să identificăm Mediana în cazul în care calculăm frecvențele cumulate. Pentru tabelul de frecvență de mai jos, mediana este acea categorie a variabilei care corespunde frecvenței cumulate care depășește 50% din observații.

În medie, câte kg de făină ați folosit în gospodăria dvs. în ultimele 6 luni?

Mediana = 3. Este valoarea variabilei care corespunde primei frecvențe cumulate care depășește 50% din cazuri (pentru valoarea 3 avem 76.25%). Valoarea anterioară (2 kg) reprezintă doar 48.72% din cazuri.

Valoarea variabilei x	**Frecvențe absolute f_i**	**Frecvențe relative (%) fr_i**	**Frecvențe cumulate**
1	120	14.62	14.62
2	280	34.10	48.72
3	226	27.53	76.25
4	98	11.94	88.19
5	52	6.33	94.52
6	45	5.48	100
	N=821		

În situația în care tabelul de frecvență cuprinde date înregistrate pe intervale, în primul rând vom determina *intervalul median* (intervalul corespunzător primei frecvențe cumulate care depășește jumătate din observații), iar apoi vom determina *mediana* după formula:

$$M_e = l + \frac{\frac{N}{2} - n_c}{f} * L$$

l- limita inferioară a intervalului median

N- numărul total de observații

nc-frecvența absolută cumulată a tuturor intervalelor care preced intervalul median

f- frecvența absolută a intervalului median

 Exemplu

Am obținut următorul tabel de frecvență pentru variabila: *Ce proporție din cunoștințele acumulate în timpul facultății consideri că vei folosi la locul de muncă?*

Valoarea variabilei pe intervale x	**Frecvențe absolute** f_i	**Frecvențe relative** fr_i	**Frecvențe cumulate**
între 0 și 10	10	0.54	0.54
între 11 și 20	31	1.67	2.22
între 21 și 30	65	3.51	5.73
între 31 și 40	102	5.51	11.24
între 41 și 50	177	9.56	20.80
între 51 și 60	206	11.13	31.93
între 61 și70	252	13.61	45.54
între **71 și 80**	**342**	**18.48**	**64.02**
între 81 și 90	358	19.34	83.36
între 91 și 100	307	16.59	100
	N=1850		

Primul pas este să identificăm intervalul median. Acesta este "între 71 și 80".
l (limita inferioară a intervalului median) este 71, iar *f* (frecvența intervalului median) este 342.
L (mărimea intervalului median) este 10.
nc = suma frecvențelor absolute ale intervalelor ce preced intervalul median
nc = 10 + 31 + 65 + 102 + 177 + 206 + 252 = 843

$$\frac{N}{2} - nc = \frac{1850}{2} - 843 = 82$$

A 82-a observație din intervalul median este Mediana.

$$\text{Mediana} = 71 + \frac{\frac{1850}{2} - 843}{342} * 10 = 71 + \frac{82}{342} * 10 = 71 + 2.40 = 73.40$$

MEDIA reprezintă suma tuturor valorilor observate ale seriei de date raportată la numărul total de indivizi statistici.

$$\bar{x} = \frac{\sum x_i * f_i}{\sum f_i}$$

x_i - valorile variabilei

f_i - frecvența absolută

Σf_i = N (numărul total)

Utilizăm simboluri diferite dacă datele sunt culese la nivelul întregii populații, sau la nivel de eșantion: $\bar{x}$ media la nivel de eșantion, respectiv μ media la nivelul populației.

ATENȚIE! Valoarea medie nu poate fi calculată decât în cazul variabilelor numerice!



Exemple

Exemplul 1. *Numărul de persoane din gospodărie*:

x	f_i
1	36
2	42
3	42
4	31
5	24
6	12
7	9
	N=196

.

$$\text{Media} = \frac{1*36 \ +2*42 \ +3*42 \ +4*31 \ +5*24 \ +6*12 \ +7*9}{196}$$

Media = 3.19

Vom citi astfel: “în medie, sunt 3.19 persoane în gospodărie” sau “numărul mediu de persoane în gospodărie este de 3.19”

Exemplul 2. *Numărul de kg de făină consumate în gospodărie în ultimele 6 luni*:

x	f_i
1	120
2	280
3	226
4	98
5	52
6	45
	N=821

Media = $\frac{1*120+2*280+3*226+4*98+5*52+6*45}{821}$ = 2.78

Vom interpreta astfel: "În medie, în 6 luni, subiecții consumă 2.78 de kg de făină" sau "Consumul mediu de făină/gospodărie într-o perioadă de 6 luni este de 2.78 de kg".

Exemplul 3. *Vârsta angajaților unei organizații:*

x	f_i
18-20	10
21-25	25
26-30	20
31-35	15
36-40	30
	N=100

Media = $\frac{19*10+23*25+28*20+33*15+38*30}{100} =$ 29.6 ani

Vom interpreta astfel: "Vârsta medie a angajaților este de 29.6 ani"

Proprietățile mediei aritmetice

- Media aritmetică este cuprinsă între valoarea minimă și cea maximă a seriei de valori ale variabilei
- Media aritmetică are aceeași unitate de măsură ca și valorile variabilei respective. De exemplu, dacă variabila vârstă se exprimă în ani și media sa aritmetică se va exprima în ani
- Suma abaterilor valorilor de la medie este nulă. Adică $\sum (x_i - \overline{x}) = 0$
- Dacă frecvențele se înmulțesc sau se împart cu același număr, valoarea mediei aritmetice nu se schimbă
- Valoarea medie poate fi calculată chiar dacă nu cunoaștem distribuția caracteristicii, ci numai suma valorilor. De exemplu dacă o organizație are n salariați și într-o lună sunt cheltuiți S lei pentru fondul de salarii (suma valorilor), atunci fără a mai urmări ce salariu (valoare) are fiecare individ, putem spune că salariul mediu este S\n
- Media aritmetică poate fi o valoare pe care nu o ia nici un individ statistic, iar uneori aceasta poate fi fără sens la nivelul indivizilor statistici. De exemplu, dacă populația statistică este formată din familiile care locuiesc într-un imobil, de exemplu în număr de 100, atunci putem construi o distribuție a acestor familii după numărul membrilor lor, aceste numere fiind valorile variabilei

Nr. persoane în gospodărie	1	2	3	4	5	Total
Nr. familii	26	32	19	14	9	100

Nu uitați! Calcularea indicatorilor tendinței centrale depinde de nivelul de măsurare! Modul poate fi determinat pentru orice tip de variabilă; Mediana poate fi determinată pentru variabilele la nivel nominal și ordinal, iar Media poate fi calculată doar pentru variabilele numerice.

Utilizarea indicatorilor tendinței centrale depinde și de existența observațiilor care au valori extreme, ceea ce se află în strânsă legătură cu forma distribuției.

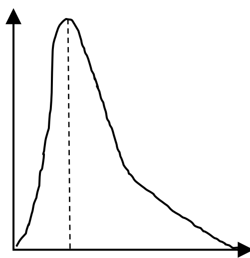
Pozițiile relative ale modului (Mo), medianei (Me) și mediei (M)

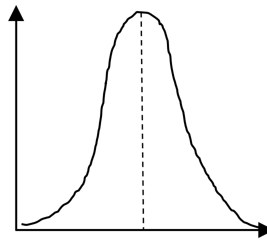
Poziționarea celor trei indicatori descrie forma distribuției.

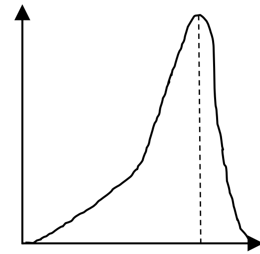
Distribuția simetrică este cea în care cei trei indicatori au aceleași valori: Mo=Me=M

Distribuția asimetrică alungită spre stânga presupune o valoare mai mică a modului față de mediană și o valoare mai mică a medianei față de medie: Mo < Me < M

Distribuția asimetrică alungită spre dreapta presupune o valoare mai mare a modului față de mediană și o valoare mai mare a medianei față de medie: M < Me < Mo







$Mo<Me<M$ $\qquad$ $Mo=Me=M$ $\qquad$ $M<Me<Mo$

Față de mod, mediana se va găsi în direcția alungirii distribuției, iar media se va găsi în aceeași direcție, chiar mai departe decât mediana.

Mediana este recomandată pentru aflarea valorii tipice când distribuția este asimetrică, fiind insensibilă la valorile extreme.

Media are avantajul că ia în considerare toate valorile seriei de date, motiv pentru care este cea mai utilizată măsură a tendinței centrale.

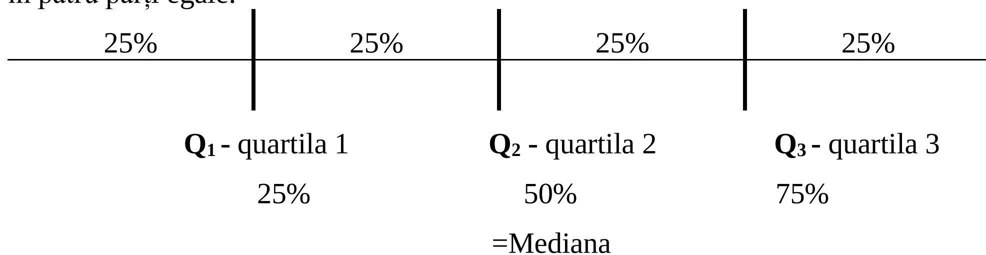
Media are dezavantajul de a fi sensibilă la valorile extreme.

Măsuri ale localizării

Măsurile localizării sunt în general cunoscute sub denumirea de percentile sau quartile.

Exemple: quartile, quintile, decile.

QUARTILELE sunt valori ale variabilei care împart mulțimea indivizilor statistici în patru părți egale.



Pentru a împărți o serie de date în patru părți egale avem nevoie de 3 valori quartile (Q_1, Q_2, Q_3). Sub Q_1 se află 25% din indivizii statistici, iar deasupra ei 75%, sub Q_2 se află 50% din indivizii statistici, iar deasupra ei 50% și sub Q_3 se află 75% din indivizii statistici, iar deasupra ei 25%. Astfel quartila 2 este tocmai mediana.

Quintilele sunt valori care împart mulțimea indivizilor statistici în cinci părți egale.

Decilele sunt valori care împart mulțimea indivizilor statistici în zece părți egale.

Pentru a împărți o serie de date în m părți avem nevoie de m-1 valori;

În statistică quartilele, quintilele, decilele se referă la valori ale variabilei.



Exemple

Variabila *Numărul de persoane din gospodărie:*

x	f_i	fr_i	f_c	
1	36	18.37	18.37	
❷	42	21.43	39.80	🡄 Q_1 (25%)
❸	42	21.43	61.22	🡄 Q_2 (50%) = Mediana
❹	31	15.82	77.04	🡄 Q_3 (75%)
5	24	12.24	89.29	
6	12	6.12	95.41	
7	9	4.59	100.00	
	N=196			

Concluzii

Indicatorii tendinței centrale sunt modul, mediana și media.

Media este cel mai precis indicator, urmat fiind de mediană și de mod.

Pentru variabile nominale, singurul indicator al tendinței centrale care poate fi calculat este modul.

Pentru variabile parametrice se folosește modul, mediana sau media.

Media nu se folosește atunci când avem valori extreme, foarte mici sau foarte mari în distribuția noastră.

Media nu poate fi folosită la date neparametrice.

Cele mai cunoscute măsuri ale localizării sunt: quartilele, quintilele și decilele.

✍ *Exerciții și aplicații*

Exercițiul 1. Se dau următoarele valori reprezentând salariile brute orare obținute de către 120 angajați dintr-o anumită ramură:

Salariu brut orar	Număr de angajați
3	20
4	35

5	30
6	20
7	15
Total	120

Determinați indicatorii tendinței centrale și valorile quartile (Q_1, Q_2, Q_3).

Exercițiul 2. Se dau următoarele valori reprezentând notele obținute de către 8 angajați ai unei organizații la examenul de finalizare a unui curs de perfecționare: 9, 7, 7, 10, 9, 9, 8, 8. Determinați indicatorii tendinței centrale

Exercițiul 3. Se dă următorul tabel:

Distribuția unui grup de indivizi statistici în funcție de variabila vârstă

Vârsta	Frecvența absolută
6 - 10	10
11 – 15	30
16 – 20	20
21 - 25	20
Total	80

Determinați intervalul modal, mediana și media.

Exercițiul 4. Am obținut următoarele răspunsuri de la subiecți cu privire la numărul de ore petrecute la calculator în medie într-o zi:

1, 2, 3, 4, 5, 6, 7, 2, 3, 1, 1, 2, 1, 3, 1, 4, 2, 1, 3, 5, 7, 1, 3, 2, 1, 2, 3, 6, 2, 1, 3, 4, 5, 2, 4, 3, 2, 1, 5, 3, 2, 1, 2, 3.

Calculați indicatorii tendinței centrale pentru această distribuție.

INDICATORI AI VARIAȚIEI (INDICATORI DE ÎMPRĂȘTIERE, INDICATORI DE DISPERSIE)

Obiective

Descrierea indicatorilor de variație.

Calcularea și interpretarea indicatorilor de variație.

Calcularea mediei și a abaterii standard pentru variabilele dihotomice

Amplitudinea, Abaterea interquartilă, Abaterea medie absolută, Varianța, Abaterea standard și Coeficientul de variație

Acești indicatori măsoară gradul de împrăștiere al indivizilor statistici în cadrul seriei de valori pe care aceștia le iau, sau cu alte cuvinte descriu omogenitatea sau eterogenitatea unei populații, (sau a unui eșantion), după caracteristicile de referință.

Există două categorii de indicatori de împrăștiere (indicatori ai variației): Indicatori (măsuri) care iau în considerare doar unele valori ale seriei de date (indicatori elementari) și indicatori (măsuri) care iau în considerare toate valorile seriei de date și le raportează la o valoare centrală (indicatori sintetici). În prima categorie sunt incluse *Amplitudinea* și *Abaterea interquartilă*, iar în ce-a de-a doua categorie sunt: *Abaterea medie absolută, Varianța*, *Abaterea standard* și *Coeficientul de variație*.

AMPLITUDINEA se bazează pe intervalul în care sunt cuprinse valorile variabilei.

$A = X_{max} - X_{min}$

X_{max} - valoarea cea mai mare a seriei de date

X_{min} - valoarea cea mai mică a seriei de date

ABATEREA INTERQUARTILĂ se referă la cele 50% din observații care sunt împrăștiate în mijlocul distribuției. În acest fel este evitată influența cazurilor extreme.

$I = Q_3 - Q_1$

Q_3 – quartila 3

Q_1 – quartila 1

 Exemplu

Variabila *Numărul de persoane din gospodărie:*

x	f_i	fr_i	f_c
1	36	18.37	18.37
2	42	21.43	39.80
3	42	21.43	61.22
4	31	15.82	77.04
5	24	12.24	89.29
6	12	6.12	95.41
7	9	4.59	100.00
	N=196		

Amplitudinea A= $X_{max} – X_{min}$ = 7 - 1= 6

Abaterea interquartilă I= Q_3-Q_1= 4 - 2= 2

Abaterea medie absolută, Varianța și Abaterea standard pornesc de la distanța valorilor variabilei față de medie. Pentru a putea calcula gradul de împrăștiere, trebuie în primul rând să observăm “îndepărtarea” fiecărei valori față de medie.

ABATEREA MEDIE este diferența dintre valoarea variabilei și media variabilei:

$x_i - \bar{x}$

Fiecare distribuție va avea un număr de abateri individuale de la medie egal cu numărul de valori ale variabilei.

ABATEREA MEDIE ABSOLUTĂ reprezintă media aritmetică a abaterilor individuale absolute de la media variabilei. Deoarece știm că una dintre proprietățile mediei constă în faptul că suma tuturor abaterilor individuale de la medie este egală cu 0, folosim valoarea absolută (ignorând semnul diferențelor)

$$\text{AM} = \frac{\sum |x_i - \bar{x}| \cdot f_i}{N}$$

VARIANȚA este tot o medie aritmetică, dar a pătratelor abaterilor individuale de la media variabilei.

La nivel de eșantion, pentru varianță utilizăm simbolul s^2 și se calculează folosind formula:

$$s^2 = \frac{\sum (x_i - \bar{x})^2 * f_i}{N-1}$$

x_i – valorile variabilei
$\bar{x}$ – media distribuției la nivel de eșantion
f_i – frecvențele înregistrate pe fiecare valoare
N – numărul total de observații

La nivelul întregii populații statistice, utilizăm simbolul σ^2 și se calculează folosind formula:

$$\sigma^2 = \frac{\sum(x_i - \mu)^2 * f_i}{N}$$

x_i – valorile variabilei
μ – media distribuției la nivel de populație
f_i – frecvențele înregistrate pe fiecare valoare
N – numărul total de observații

ABATEREA STANDARD radical de ordinul doi (rădăcină pătrată) din varianță.

La nivel de eșantion folosim simbolul *s*:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum(x_i - \bar{x})^2 * f_i}{N-1}}$$

La nivelul populației statistice:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum(x_i - \mu)^2 * f_i}{N}}$$

Media și abaterea standard au aceeași unitate de măsură; cu cât abaterea standard este mai mare în raport cu media, cu atât populația (sau eșantionul) este mai eterogenă, și cu cât abaterea standard este mai mică în raport cu media cu atât populația (eșantionul) este mai omogenă.

Media și abaterea standard sunt cei mai des utilizați indicatori statistici.

NU UITAȚI: acești indicatori pot fi calculați doar pentru variabilele numerice (cantitative).

 Exemple

Am obținut pe un eșantion de 130 de studenți, răspunsurile la întrebarea *În medie, câte ore pe zi alocați pregătirii seminariilor, cursurilor și laboratoarelor?*
Tabelul de frecvență cuprinde următoarele date:

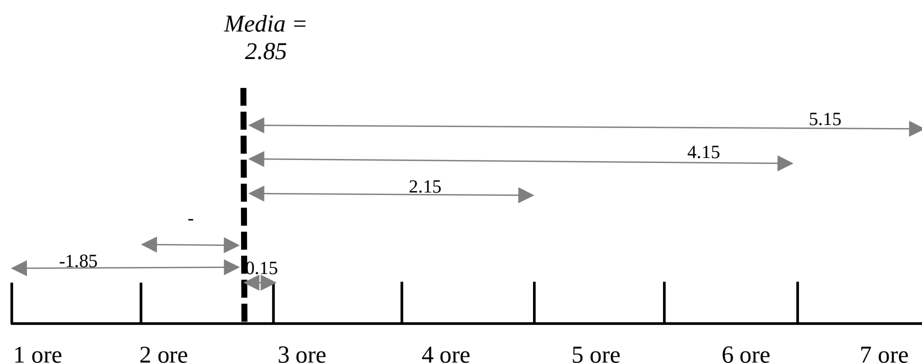
x (ore)	f_i
1	31
2	45
3	28
5	12
7	6
8	8
	N=130

Media este $\bar{x} = \frac{\sum x_i * f_i}{N} = \frac{1*31+2*45+3*28+5*12+7*6+8*8}{130} = 2.85$

În medie, studenții dedică 2.85 ore pentru pregătirea seminariilor și cursurilor. Dar nu este suficient să știm acest lucru. Ne interesează cât de diferiți sunt studenții între ei? Cât de mult se "abat" ei de la această medie? Dacă sunt mulți studenți

care se "îndepărtează" de la această medie (adică dacă obținem o abatere standard cu o valoare ridicată), putem considera că avem o populație eterogenă: sunt studenți care se pregătesc pentru cursuri 2-3 ore, dar sunt și studenți diferiți de aceștia, care, să presupunem, dedică 10-12 ore pentru pregătirea cursurilor și seminariilor. Ca să aflăm acest lucru, pornim de la abaterile individuale de la medie $x_i - \bar{x}$.

x	Abaterile individuale de la medie $x_i - \bar{x}$
1	1 - 2.85 = -1.85
2	2 - 2.85 = -0.85
3	3 - 2.85 = 0.15
5	5 - 2.85 = 2.15
7	7 - 2.85 = 4.15
8	8 - 2.85 = 5.15



Din imagine, dar și observând valoarea de 5.15, înțelegem că studenții care au răspuns că dedică 8 ore pentru pregătirea cursurilor se poziționează cel mai departe față de medie. În continuare, dorim să aflăm, în general studenții cât se abat de la această valoare medie de 2.85 ore.

Varianța o vom calcula ca media aritmetică a pătratelor acestor abateri individuale de la media distribuției:

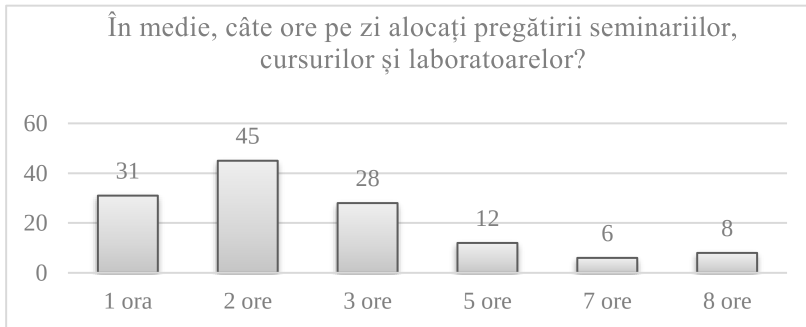
$$s^2 = \frac{\sum (x_i - \bar{x})^2 * f_i}{N - 1}$$

$$s^2 = \frac{(1-2.85)^2*31+(2-2.85)^2*45+(3-2.85)^2*28+(5-2.85)^2*12+(7-2.85)^2*6+(8-2.85)^2*8}{130-1}$$

$$s^2 = \frac{510.22}{129} = 3.96$$

Abaterea standard este radical din varianță:
$s = \sqrt{s^2} = \sqrt{3.96} = 1.99$

Ne vom exprima astfel: "În medie, studenții dedică 2.85 ore pentru pregătirea cursurilor și seminariilor, cu o abatere standard de 1.99, ceea ce înseamnă că, în medie, cei 130 de studenți din eșantion se abat de la media de 2.85 ore cu 1.99 ore".





Exemplu

În cadrul unui sondaj am adresa următoarea întrebare subiecților din eșantion: *Într-o zi din cursul săptămânii, în medie, câte ore petreceți vizionând programe la TV?*, și am obținut următoarele răspunsuri: 1, 2, 4, 1, 3, 4, 5, 2, 1, 3, 2, 4, 1, 3, 5, 4, 5, 3, 4, 1, 2, 3, 2, 2, 2, 3, 1, 2, 3, 4, 2, 3, 2, 2, 3.

Pentru a realiza tabelul de frecvență, vom identifica valorile variabilei: 1, 2, 3, 4, 5, și vom identifica frecvența pentru fiecare valoare în parte:

x (ore)	f_i
1	6
2	11
3	9
4	6
5	3
Total (N)	**35**

Media aritmetică va fi:

$$\bar{x} = \frac{\sum x_i * f_i}{N} = \frac{1*6+2*11+3*9+4*6+5*3}{35} = 2.69$$

Abaterea standard o vom calcula astfel:

$$s^2 = \frac{(1-2.69)^2 * 6 + (2-2.69)^2 * 11 + (3-2.69)^2 * 9 + (4-2.69)^2 * 6 + (5-2.69)^2 * 3}{35-1}$$

$s^2 = 1.21$

În medie, subiecții din eșantion petrec 2.69 ore la TV într-o zi din cursul săptămânii, cu o abatere standard de 1.21.

COEFICIENTUL DE VARIAȚIE este o măsură adimensională. El este foarte util în compararea variației a două variabile măsurate pe aceeași populație/eșantion. Poate fi determinat numai pentru variabile măsurate la nivel de raport.

La nivel de eșantion:

$$CV = \frac{s}{\bar{x}}$$

La nivelul populației statistice:

$$CV = \frac{s}{\mu}$$

Pentru o interpretare mai ușoară poate fi înmulțit cu 100 și indică procentul (%) din medie ce corespunde unei abateri standard.

 Exemplu

Pentru datele din exemplul anterior,

$$CV = \frac{1.99}{2.85} = 0.70$$

70% din medie corespunde unei abateri standard.

Calcularea mediei și a abaterii standard pentru variabilele dihotomice

Media unei variabile dihotomice este chiar frecvența relativă de apariție a valorii "1".

 Exemplu

Am obținut următoarele rezultate cu privire la intenția angajaților de a participa la un curs de calificare:

Intenția de a participa la un curs de calificare	Frecvențe absolute f_i	Frecvențe cumulate f_{ri}
DA → "1"	1200	0.8
NU → "0"	300	0.2
Total	1500	

Codul 1 indică prezența caracteristicii, iar codul 0 absența caracteristicii.
Media o vom calcula astfel:
$\bar{x} = \frac{\sum x_i * f_i}{\sum f_i} = \frac{(1*1200)+(0*300)}{1200+300} = \frac{1200}{1500} = 0.8$ Această valoare o vom nota cu p
Media unei variabile dihotomice este frecvența relativă de apariție a valorii "1".
Abaterea standard o vom calcula folosind formula:
$s = \sqrt{s^2} = \sqrt{p(1-p)} = \sqrt{0.8(1-0.8)} = 0.4$

Concluzii

Indicatorii de împrăștiere pot lua în considerare doar unele valori ale seriei de date și se numesc indicatori elementari ai împrăștierii sau toate valorile seriei de date, cazul indicatorilor sintetici

Indicatorii elementari ai împrăștierii sunt amplitudinea variației și abaterea interquartilă, iar indicatorii sintetici cei mai utilizați sunt: abaterea medie absolută, varianța, abaterea standard și coeficientul de variație.

✍ *Exerciții și aplicații*

Exercițiul 1. Se dau următoarele valori reprezentând notele obținute de către 8 elevi dintr-un Centru de Educație Incluzivă, la un test de matematică:

6, 5, 8, 7, 6, 5, 9, 6

Calculați indicatorii de împrăștiere.

Exercițiul 2. Se dau următoarele valori reprezentând numărul de copii din gospodărie dintr-un imobil:

1, 3, 2, 3, 1, 4, 5, 2, 4, 2.

Calculați varianța și abaterea standard.

Exercițiul 3. Se dau următoarele valori, reprezentând salariile brute orare, exprimate în euro obținute de către zece angajați ai unei organizații:

Angajatul A	3
Angajatul B	2.5
Angajatul C	4
Angajatul D	2.5
Angajatul E	5
Angajatul F	3
Angajatul G	4
Angajatul H	3
Angajatul I	3
Angajatul J	2.5

Determinați: Frecvențele relative, Indicatorii tendinței centrale, Abaterea medie absolută și Abaterea standard.

Exercițiul 4. Determinați media și abaterea standard pentru variabila dihotomică din tabelul de mai jos:

Intenția de a susține candidatul „A” pentru funcția de director	Frecvențe absolute
DA	400
NU	300
Total	700

PROBABILITĂȚI. DISTRIBUȚIA NORMALĂ

Obiective

Înțelegerea conceptului de probabilitate

Importanța distribuției normale

Înțelegerea proprietăților distribuției normale

Calcularea scorului z

Citirea corectă a tabelului z

Ce reprezintă probabilitatea

Multe dintre evenimente nu sunt predictibile. Atunci când vorbim de probabilitatea ca un eveniment să aibă loc, ne referim la șansele ca acesta să se întâmple.

Probabilitatea unui eveniment este chiar frecvența relativă de apariție a acelui eveniment.

 Exemple

Șansa ca o monedă să cadă pe fața cu banul este de 1/2, deoarece este o singură față cu ban din două fețe posibile. Cu alte cuvinte, probabilitatea ca moneda să cadă pe fața cu banul este de 1/2 = 0.5.
Probabilitatea ca, aruncând un zar (care are 6 cifre), să obținem cifra 3 este de 1/6 = 0.17, deoarece este o singură față cu cifra 3 din 6 posibile.
Probabilitatea de a alege la întâmplare un băiat dintr-o grupă de 100 de elevi, în care 70 sunt fete și 30 sunt băieți, este de P(B)=30/100=0.3 (sunt 30 de șanse să fie băiat din 100 de posibilități în total). Iar probabilitatea de a alege o fată este de P(F)=70/100=0.7.
Dacă avem o pungă cu 72 de bomboane, 14 cu aromă de căpșuni, 25 cu aromă de zmeură și 33 cu aromă de cireșe, șansa ca la întâmplare să alegem o bomboană cu aromă de zmeură este de P(zmeură)=25/72 = 0.35.

Orice probabilitate variază între valorile 0 și 1. O probabilitate apropiată de valoarea 0 indică șansele foarte reduse ca evenimentul să se întâmple, iar o probabilitate de 1 exprimă siguranța ca evenimentul să aibă loc.

Spunem că două evenimente, A și B, sunt independente dacă P(A și B)= P(A)*P(B).

Distribuția normală

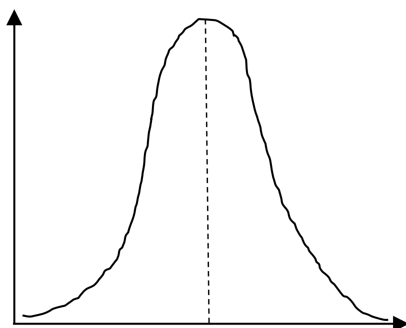
Distribuția normală permite aproximarea distribuțiilor multor variabile întâlnite în aplicațiile statistice din științele sociale și este utilizată în multe metode de inferență statistică. Astfel, dacă avem date culese la nivelul unui eșantion, putem să estimăm anumite valori ale variabilelor, cunoscând cu precizie care sunt șansele să greșim (eroarea estimării).

O distribuție normală este pe deplin caracterizată de medie, ca indicator al tendinței centrale și de abatere standard, ca indicator al variației. Acești doi indicatori se numesc **parametri ai distribuției normale**. Dacă cunoaștem media și abaterea standard, putem calcula probabilitatea unei valori particulare în această distribuție.

Pentru statisticile la nivel de eșantion și cele la nivel de populație sunt utilizate notări diferite.

	La nivelul populației statistice	La nivel de eșantion
media	μ	$\overline{x}$
abaterea standard	σ	s
frecvența relativă	π	p

Distribuția normală a fost descrisă prima dată de Ch. Fr. Gauss, și din acest motiv distribuția normală se mai numește și distribuție gaussiană. Întrucât la demonstrarea acestui concept a participat și P.S. Laplace, în literatura de specialitate se va întâlni și termenul de distribuție gauss – laplace. Datorită formei distribuției (asemănarea cu forma unui clopot), distribuția normală este cunoscută și sub denumirea "clopotul lui Gauss". Forma de clopot a distribuției este determinată de valoarea abaterii standard față de centrul dat de valoarea medie.



Principala proprietate a distribuției normale este simetria: Mo = Me=M.

Pentru ca o distribuție să fie considerată normală, vor trebui îndeplinite simultan următoarele condiții:

- să fie unimodală – adică să existe un singur mod, o singură categorie cu frecvență maximă
- să fie simetrică față de medie – adică să nu fie deplasată spre stânga sau spre dreapta
- să fie normal boltită – adică să nu fie nici ascuțită (foarte omogenă) și nici turtită (foarte eterogenă).

Ariile de sub curba normală reprezintă probabilități (aria totală de sub curba normală este 100% sau probabilitate 1). Deoarece distribuția este simetrică, aria unei jumătăți este egală cu 50% (sau 0.5). 50% din observații se află în stânga mediei, 50% se află în dreapta ei.

Una dintre cele mai importante proprietăți ale distribuției normale se referă la faptul că pentru orice număr **z**, probabilitatea regăsită la dreapta valorii μ+zσ este aceeași pentru orice distribuție normală. Astfel, aria creionată de curba normală în dreapta valorii μ+zσ este dată doar de valoarea lui **z**.

μ+zσ reprezintă probabilitatea apariției unei valori la ***z*** abateri standard în dreapta mediei. Distanța dintre medie și o valoare (x_i) se măsoară în abateri standard. Deoarece proprietățile distribuției normale sunt aceleași, indiferent de valoarea mediei și a abaterii standard, pentru a determina aria dintre medie și o valoare (x_i) vom apela la un caz special al distribuției normale - *distribuția normală normată sau standard* care are μ=0 și σ=1. Aceasta o regăsim și cu denumirea Tabel z (valorile lui z sunt prezentate în Anexa 1). Scorurile z indică îndepărtarea unei valori față de media unei distribuții, în abateri standard față de medie.

Pentru o distribuție normală, probabilitatea ca o variabilă să ia valori între

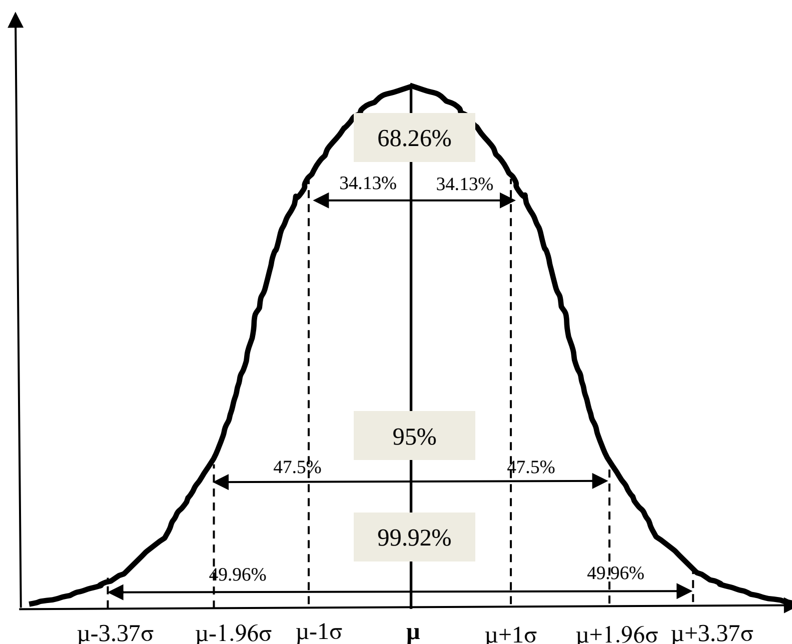
μ-1.96σ și μ+1.96σ este de 0.95. Sunt 95% șanse ca valoarea să se afle în acest interval. Cu alte cuvinte, 95% din cazuri se află la 1.96 abateri standard față de medie.

47.5% din cazuri sunt în dreapta mediei, iar 47.5% sunt în stânga mediei. Astfel, în total, intervalul (μ-1.96σ; μ+1.96σ) cuprinde 95% din cazuri.

Dacă z=1 aria dintre μ-zσ și μ+zσ este 0.682. 68.26% din cazuri sunt distribuite la 1 abatere standard față de medie. 34.13% sunt în stânga mediei, iar 34.13% sunt în dreapta mediei.

Dacă z=2.58 aria dintre μ-zσ și μ+zσ este 0.99. 49.51% dintre valori sunt în stânga mediei, iar 49.51% sunt în dreapta. În total, aria acoperă 99% din valori.

Dacă z = 3.37, aria dintre μ-zσ și μ+zσ este 0.999. 99.92% dintre valori se află la 3.37 abateri standard față de medie (49.96% sunt în stânga mediei, iar 49.96% sunt în dreapta).



Toate aceste valori calculate pentru fiecare nivel de probabilitate sunt cuprinse în tabele statistice.

Procedura prin care se determină la câte abateri standard față de medie se găsește o valoare x_i se numește *standardizare*.

$$z_{x_i} = \frac{|x_i - \mu|}{\sigma}$$

 Exemple

Exemplul 1. Să presupunem că media coeficientului de inteligență al unei populații de studenți se distribuie normal în jurul unei medii de 135 ($\mu = 135$) și cu o abatere standard de 12 ($\sigma = 12$). Care este probabilitatea ca selectând un student la întâmplare acesta să aibă coeficientul de inteligență mai mic de 126?

Pentru a răspunde la această întrebare trebuie să parcurgem următoarele etape:
1. determinăm la câte abateri standard față de medie (135) se găsește valoarea 126, practic standardizăm valoarea 126.

Convertirea valorii 126 în scor z înseamnă o transformare a distribuției normale inițiale (cu media 135 și abaterea standard 12) într-o distribuție normală normată sau standard (cu media 0 și abaterea standard 1)

$z_{126} = \frac{|126-135|}{12} = 0.75$, de unde rezultă că 126 se găsește la 0.75 abateri standard față de 135 (medie).

2. Cu ajutorul tabelului z determinăm aria dintre 126 și 135. Având în vedere că z=0.75 ne vom uita la intersecția rândului ”0,7” cu coloana ”0.05” (adunate cele 2 valori dau exact 0.75) și vom găsi valoarea 0.2734

3. Determinarea probabilității ca selectând un student la întâmplare acesta să aibă coeficientul de inteligență mai mic de 126 înseamnă de fapt determinarea ariei de la stânga lui 126. Pentru a determina aria de la stânga valorii 126 ne vom folosi de proprietatea distribuției normale care ne spune că aria totală a distribuției normale este 1, distribuția este simetrică, ceea ce înseamnă că aria unei jumătăți este 0,5. Astfel, vom scădea din 0.5 valoarea de 0.2734 și vom obține 0.2266, ceea ce înseamnă că probabilitatea ca selectând un student la întâmplare acesta să aibă coeficientul de inteligență mai mic decât 126 este de 22.66% (0.2266x100)

Exemplul 2. Să presupunem că rezultatele (notate cu valori cuprinse între 0 și 100) obținute la un test de statistică de către studenții Facultății de Științe Socio-Umane se distribuie normal în jurul unei medii de 70 puncte și cu o abatere standard de 5 puncte. Care este probabilitatea ca selectând un student la întâmplare acesta să aibă un rezultat cuprins între 75 și 79 puncte?

Pentru a răspunde la această întrebare trebuie să parcurgem următoarele etape:

1. determinăm la câte abateri standard față de medie (70) se găsesc valorile 75 și respectiv 79, practic standardizăm cele două valori.

Convertirea valorilor 75 și 79 în scoruri z înseamnă o transformare a distribuției normale inițiale (cu media 70 și abaterea standard 5) într-o distribuție normală normată sau standard (cu media 0 și abaterea standard 1)

$z_{75} = \frac{|75-70|}{5} = 1$ de unde rezultă că 75 se găsește la 1 abatere standard față de 70 (medie)

$z_{79} = \frac{|79-70|}{5} = 1.8$ de unde rezultă că 79 se găsește la 1.8 abateri standard față de 70 (medie).

2. Cu ajutorul tabelului z determinăm aria dintre 0 și 1 și aria dintre 0 și 1.8. Conform tabelului z acestea sunt 0.3413 și respectiv 0.4641.

3. Determinarea probabilității ca selectând un student la întâmplare acesta să aibă un punctaj cuprins între 75 și 79 înseamnă de fapt determinarea ariei dintre cele 2 valori. Astfel, vom efectua scăderea 0.4641- 0.3413 și vom obține rezultatul 0.1228, ceea ce înseamnă că probabilitatea c ca selectând un student la întâmplare acesta să aibă un punctaj cuprins între 75 și 79 este de 12.28%.

Concluzii

Probabilitatea unui eveniment este chiar frecvența relativă de apariție a acelui eveniment.

Media și abaterea standard caracterizează pe deplin o distribuție, din acest motiv cei doi indicatori poartă numele de parametri ai distribuției normale

Scorul z reprezintă distanța dintre medie și o valoare (x_i) în abateri standard.

Pentru determinarea ariei dintre medie și o valoare (x_i) se apelează la un caz special al distribuției normale, care se numește distribuția normală normată sau standard care are μ=0 și σ=1 (Tabel z).

✍ *Exerciții și aplicații*

Exercițiul 1. În cadrul unei facultăți structura personalului didactic este următoarea: 15 asistenți, 34 lectori, 12 conferențiari și 5 profesori. Determinați:

a) Care este probabilitatea ca selectând un cadru didactic la întâmplare acesta să fie asistent.

b) Care este probabilitatea ca selectând un cadru didactic la întâmplare acesta să fie conferențiar.

c) Care este probabilitatea ca selectând un cadru didactic la întâmplare acesta să fie profesor.

Exercițiul 2. Vârsta femeilor angajate într-o organizație care oferă servicii sociale se distribuie normal în jurul unei medii de 45.31 ani și cu o abatere standard de 15.76 ani. Care este probabilitatea ca selectând o persoană la întâmplare aceasta să aibă vârsta mai mică de 30 ani?

Exercițiul 3. Să presupunem că în cadrul unei populații statistice variabila greutate se distribuie normal în jurul unei medii de 69 kg și cu o abatere standard de 1.6 kg. a) Care este probabilitatea ca selectând un individ statistic la întâmplare acesta să aibă o greutate cuprinsă între 67 și 71 kg?

b) Care este probabilitatea ca selectând un individ statistic la întâmplare acesta să aibă o greutate mai mică de 67 kg?

c) Care este probabilitatea ca selectând un individ statistic la întâmplare acesta să aibă o greutate mai mare de 71 kg?

Exercițiul 4. Vârsta unei colectivități de persoane cu drept de vot se distribuie normal în jurul unei medii de 47 ani și cu o abatere standard de 16.6 ani. Care este probabilitatea ca selectând o persoană la întâmplare aceasta să aibă vârsta mai mică de 50 ani?

EȘANTIONAREA

Obiective

Înțelegerea semnificației conceptelor: eșantionare, distribuție de eșantionare, eroare de eșantionare, intervale de încredere

Calculul erorii standard, al erorii de eșantionare și a intervalelor de încredere pentru medie și proporții

Cunoașterea conceptului de reprezentativitate și a factorilor care influențează gradul de reprezentativitate a eșantionului

Cunoașterea procedurilor de eșantionare

Problematica cercetărilor pe eșantioane

În cele mai multe cazuri, cercetările sociologice necesită investigarea unei populații de dimensiuni foarte mari. În asemenea situații, cercetătorii se află în imposibilitatea de a cuprinde toți indivizii în cadrul cercetării (din cauza consumului de resurse și timp) și astfel este necesară selecția doar unei anumite părți din indivizii cuprinși în universul cercetării. Problema care apare în acest context se referă la încrederea pe care o poate avea cercetătorul că datele obținute doar de la o parte din indivizi reflectă într-o bună măsură "realitatea" corespunzătoarele întregului univers al cercetării. Cu alte cuvinte, eliminând din studiu anumiți indivizi, care este probabilitatea ca rezultatele obținute să se îndepărteze față de valorile pe care le-ar fi obținut dacă ar fi inclus toți indivizii în cercetare. Poate cercetătorul să generalizeze date obținute de la un lot de subiecți la nivelul întregii populații? Din punct de vedere statistic putem să identificăm această "probabilitate de a greși" , dar este extrem de important să ținem cont de anumite condiții, iar una dintre acestea se referă la modul în care au fost selectați indivizii (procedura de eșantionare).

Eșantionarea se referă la metodele sistematice de selecție a subiecților care urmează să fie cuprinși în cadrul studiului. Prin culegerea datelor la nivel de eșantion putem să cunoaștem întreaga populație din care a fost selectat eșantionul, prin realizarea de inferențe. Totuși, în această situație trebuie să cunoaștem probabilitatea de a face erori de predicție. În funcție de modul în care au fost selectați indivizii cuprinși în eșantion, aceste erori de predicție pot fi calculate.

Potrivit teoriilor matematice (legile numerelor mari), dacă sunt respectate condițiile de independență, pe măsură ce repetăm un experiment (o selecție aleatoare a unui individ dintr-o mulțime), vom observa că distribuția rezultatelor obținute sunt din ce în ce mai asemănătoare; distribuția obținută se apropie din ce în ce mai mult de modul în care s-ar distribui toate rezultatele situațiilor posibile. Cu alte cuvinte, dacă dorim să măsurăm o anumită variabilă la nivelul unei populații (de exemplu înălțimea), este suficient să măsurăm această variabilă de la indivizi cuprinși în

populație, cu condiția ca selecția indivizilor să fie independentă de celelalte selecții posibile. Teoriile matematice ne spun că pe măsură ce vom selecta indivizii, distribuția rezultatelor variabilei măsurate se va apropia de distribuția variabilei la nivelul populației. Ca atare, dacă vom selecta corect (aleator) indivizii cuprinși în eșantion, rezultatul obținut în urma măsurării va fi foarte asemănător cu cel pe care l-am fi obținut dacă obțineam datele la nivelul întregii populații. Probabilitatea să greșim atunci când facem inferența va fi mică.

Atunci când lucrăm cu un eșantion, trebuie să înțelegem că vom avea trei tipuri de distribuții a unei variabile: distribuția variabilei la nivelul eșantionului (de mărime n), distribuția variabilei la nivelul întregii populații (de mărime N) și distribuția mediilor variabilei obținute pe toate eșantioanele posibile de mărime n extrase din populație. Putem să ne imaginăm că valorile medii ale variabilei obținute pe toate eșantioanele posibile formează ele în sine o variabilă (care se va distribui într-un anumit fel, va avea o valoare medie și o abatere standard). Aceasta distribuție poartă denumirea de ***distribuția de eșantionare a mediei***. Dacă numărul indivizilor selectați este suficient de mare, atunci distribuția de eșantionare a mediei (distribuția mediilor obținute pe toate eșantioanele posibile) va fi o distribuție normală, ca atare va avea toate proprietățile unei distribuții normale.

Potrivit *Teoriei Limitei Centrale*, această distribuție de eșantionare a mediei (distribuție normală) va avea media aceeași cu cea regăsită la nivelul populației (μ) și va avea o abatere standard calculată ca

$e = \frac{\sigma}{\sqrt{n}}$ unde σ este abaterea standard din populație, iar n este mărimea eșantionului

Această abatere standard este notată cu e și este numită ***eroare standard.***

Vom sintetiza cele menționate mai sus: Să presupunem că se extrag eșantioane repetate din aceeași populație (de exemplu 300.000 de persoane). Câte eșantioane de 1000 de persoane se pot extrage din 300.000? Combinații de 300.000 luate câte 1000.

- Pentru toate aceste eșantioane vom calcula media parametrului estimat (de exemplu înălțimea)
- Dacă avem un număr suficient de mare de eșantioane, toate valorile medii ale variabilei (toate mediile înălțimilor obținute pe toate eșantioanele) se vor distribui sub forma unei distribuții normale (clopotul lui Gauss). Aceasta este distribuția de eșantionare a mediei pe eșantioane.
- Cu cât numărul eșantioanelor este mai mare, cu atât mai mult se apropie distribuția de rigorile unei distribuții normale.
- Această distribuție va avea media egală cu media din populație și abaterea standard (eroarea standard) calculată ca abaterea standard a variabilei la nivelul populației, raportată la radical din dimensiunea eșantionului.

Ținând cont de caracteristicile distribuției normale (prezentate în capitolul anterior) putem să calculăm pentru fiecare valoare măsurată la nivel de eșantion, probabilitatea ca valoarea regăsită la nivelul populației să se încadreze într-un anumit interval față de valoarea obținută pe eșantion. Acest interval este denumit ***interval de încredere*** sau ***interval de confidență.***

Intervale de încredere (de confidență)

Rezultatele obținute în cadrul unei cercetări selective (pe baza unui eșantion) sunt afectate de erori, iar ceea ce se poate obține printr-o astfel de cercetare nu este adevărata valoare din populație (parametru), ci o estimare a acesteia. În această situație, în funcție de modul în care a fost construit eșantionul, putem să calculăm cât de mult greșim când facem această estimare. Cu alte cuvinte, care este probabilitatea ca estimarea să fie una de încredere. În cadrul cercetării ne dorim să avem o eroare "suficient" de mică și o probabilitate "suficient" de mare ca datele obținute pe eșantion să corespundă cu cele regăsite la nivelul întregii populații. Vorbim în acest sens de reprezentativitatea eșantionului și anume capacitatea acestuia de a reproduce cât mai corect caracteristicile populației. Gradul de reprezentativitate este exprimat cantitativ ținând cont de eroarea maximă și de nivelul de probabilitate, aspecte prezentate în capitolul de față.

Pornind de la valoarea obținută la nivelul eșantionului, cu o anumită probabilitate dată de către cercetător, poate fi calculat un interval de încredere (de confidență) în care se încadrează valoarea obținută la nivelul populației. Cel mai mic nivel de probabilitate (de încredere) acceptat în științele sociale este de 95% (0.95), ceea ce înseamnă că șansele de a greși estimarea nu trebuie să fie mai mari de 5%. Trebuie să ne asigurăm că eroarea de eșantionare este sub limita acceptată ("greșim mai puțin decât ne este permis") și avem o încredere de peste 95% că estimările sunt corecte.

Pentru nivelul de încredere de 95%, vom avea:

P = 95%	P = 0.95
p = 100% - P = 100% - 95% = 5%	p = 1 – P = 1 – 0.95 = 0.05

Limitele intervalului de încredere (de confidență) se calculează cu ajutorul valorii lui z corespunzătoare nivelului de probabilitate (P) și a erorii de eșantionare specifice variabilei măsurate, adică a erorii standard (abaterea standard pentru distribuția de eșantionare a mediei).

Cu o probabilitate P putem spune că valoarea variabilei la nivelul populației statistice se încadrează în intervalul:

(valoarea variabilei pe eșantion – z*e; valoarea variabilei pe eșantion + z*e)

unde $e = \frac{\sigma}{\sqrt{n}}$**, iar pentru proporții (frecvențe relative)** $e = \sqrt{\frac{p(1-p)}{n}}$

e este eroare standard;

σ este abaterea standard la nivelul populației statistice; deoarece de cele mai multe ori parametrii sunt necunoscuți, iar σ repezintă un parametru, aceasta este înlocuită cu s (abaterea standard la nivel de eșantion) întrucât aceasta reprezintă cea mai bună estimare pentru σ;

n este mărimea eșantionului;

p este proporția sau frecvența relativă la nivel de eșantion.

Identificarea intervalului de încredere pentru medie

$P(\bar{x} - \Delta x < \mu < \bar{x} + \Delta x)$ = N.î.(nivel de încredere)

$\mu \in [\bar{x} - \Delta_x ; \bar{x} + \Delta_x]$ Vom citi astfel: media la nivelul populației (μ) este cuprinsă în intervalul dintre media obținută pe eșantion ($\bar{x}$) $\pm$ eroarea de eșantionare (Δ_x).

$\Delta_x = z^*e$ unde z este coeficient tabelar, iar e eroare standard.

Astfel, $\mu \in [\bar{x} - z^*e ; \bar{x} + z^*e]$

$$\mu \in [\bar{x} - z^*\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}} ; \bar{x} + z^*\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}]$$

Pentru nivelul de probabilitate de 95% (P= 95%; p=0.05) valoarea coeficientului tabelar z este 1.96 (z =1.96). Astfel, eroarea de eșantionare este Δ_x=1.96*e.

Forma generală a intervalului de încredere pentru medie corespunzătoare nivelului de probabilitate de 95% poate fi scrisă astfel:

$P(x - 1.96e < \mu < x + 1.96e) = 95\%$,

iar media variabilei la nivelul populației statistice se găsește în intervalul

$\mu \in [\bar{x} - 1.96e; \bar{x} + 1.96e]$, și anume

$\mu \in [\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} ; \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}]$

Există 95% șanse ca valoarea obținută pe eșantion să se afle la o distanță de 1.96 erori standard față de valoarea medie la nivelul populației.

Pentru nivelul de probabilitate de 99% (P= 99%) valoarea coeficientului tabelar z este 2.58 (z=2.58), iar eroarea de eșantionare Δ_x=2.58*e

Media variabilei la nivelul populației statistice se găsește în intervalul

$\mu \in [\bar{x} - 2.58e; \bar{x} + 2.58e]$, și anume

$\mu \in [\bar{x} - 2.58\frac{\sigma}{\sqrt{n}}; \bar{x} + 2.58\frac{\sigma}{\sqrt{n}}]$

Identificarea intervalului de încredere pentru proporții

$P(p - \Delta_p < \pi < p + \Delta_p) = N.î$

$\pi \in [p - \Delta_p; \ p + \Delta_p]$. Vom citi astfel: valoarea la nivelul populației (π) se încadrează în intervalul dintre valoarea obținută la nivelul eșantionului (p) ± eroarea de eșantionare (Δ_p).

Pentru nivelul de probabilitate de 95%, vom avea $\Delta_p = z*e = 1.96e$

Valoarea la nivelul populației se va încadra în intervalul

$\pi \in [p - 1.96e; \ p + 1.96e]$, și anume

$\pi \in [p - 1.96\sqrt{\frac{p(1-p)}{n}}; \ p + 1.96\sqrt{\frac{p(1-p)}{n}}]$

Pentru nivelul de probabilitate de 99%, vom avea $\Delta_p = z*e = 2.58e$

Valoarea la nivelul populației se va încadra în intervalul

$\pi \in [p - 2.58e; \ p + 2.58e]$, și anume

$\pi \in [p - 2.58\sqrt{\frac{p(1-p)}{n}}; \ p + 2.58\sqrt{\frac{p(1-p)}{n}}]$

 Exemple

Exercițiul 1. Pe un eșantion aleatoriu de 1600 de elevi din clasa XI-XII, am obținut 2.7 ore valoarea medie cu privire timpul alocat pregătirii lecțiilor, cu o abatere standard de 1.2. Care este intervalul de încredere în care se încadrează valoarea la nivelul populației pentru un prag de încredere de 95%?
n = 1600 elevi
$\bar{x}$ = 2.7 ore
σ = 1.2
Media la nivelul populației se încadrează în următorul interval
$\mu \in [\bar{x} - z*e ; \bar{x} + z*e]$ unde z pentru pragul de 95% este 1.96, iar
$e = \frac{\sigma}{\sqrt{n}} = \frac{1.2}{\sqrt{1600}} = \frac{1.2}{40} = 0.03$
$\mu \in [2.7 - 1.96*0.03; 2.7 + 1.96*0.03]$
$\mu \in [2.7 - 0.059; 2.7 + 0.059]$
$\mu \in [2.64; 2.76]$
Sunt 95% șanse ca media numărului de ore dedicat pregătirii lecțiilor la nivelul populației de elevi să ia una dintre valorile din intervalul cuprins între 2.64 și 2.76.

Exercițiul 2. Pe un eșantion de 1400 de subiecți am obținut 27% procentul celor care declară că au obiceiul zilnic de a lua gustări, în afara meselor principale, în timp ce se uită la TV. Pentru pragul de încredere de 95%, care este valoarea la nivelul populației:
n = 1400
p = 0.27 (27%)

z = 1.96 pentru P = 95%
procentul la nivelul populației se va încadra în intervalul
$\pi \in [\, p - z*e;\ p + z*e]$, unde $e = \sqrt{\frac{p(1-p)}{n}}$

$$e = \sqrt{\frac{0.27(1-0.27)}{1400}} = \sqrt{\frac{0.27*0.73}{1400}} \sqrt{\frac{0.197}{1400}} = \sqrt{0.000141} = 0.0119 = 1.19\%$$

Intervalul este [27% – 1.96*1.19; 27% +1.96*1.19]
[27% – 2.33; 27% +2.33]
[24.67; 29.33]
Sunt 95% șanse ca procentul la nivelul populației să ia valori între 24.67% și 29.33%.

Aspecte privind reprezentativitatea

Reprezentativitatea se referă la posibilitatea ca valorile obținute la nivel de eșantion să reproducă cât mai corect și fidel structura și caracteristicile populației din care este extras. Reprezentativitatea eșantionului presupune o aproximare cât mai corectă a caracteristicilor populației din care a fost selectat eșantionul pe baza caracteristicilor populației din eșantion.

Un eșantion nu este reprezentativ în general, ci are o anumită reprezentativitate pentru fiecare caracteristică în parte. De exemplu, un eșantion de studenți are o anumită reprezentativitate în raport cu vârsta și altă reprezentativitate în raport rezultatele obținute la un examen. Pentru un eșantion se va stabili pe de o parte dacă acesta este sau nu reprezentativ, iar pe de altă parte este important să știm cât de reprezentativ este în funcție de caracteristica populației avută în vedere.

A măsura reprezentativitatea presupune a aprecia măsura în care estimările din eșantion se abat de la parametri din populație. Dar, de regulă, parametri din populație nu sunt cunoscuți, altfel nu are sens realizarea cercetării pe un eșantion. În această situație putem înlocui eroarea standard a distribuției de eșantionare cu eroare standard a eșantionului.

Reprezentativitatea se exprimă cantitativ prin:

eroarea de eșantionare (eroarea limită de sondaj) $\Delta_x = z*e$

Erorile de eșantionare (aleatoare) trebuie deosebite de cele sistematice, datorate unor erori în realizarea cercetării. Cele aleatoare nu sunt datorate unor greșeli ale cercetătorului, ci variabilității eșantionului selectat din populație. Evident, cu cât eșantionul este mai mare, cu atât erorile de eșantionare vor fi mai mici. Însă, cu creșterea mărimii eșantionului, crește și probabilitatea de apariție a erorilor sistematice (nealeatoare).

și *nivelul de încredere/probabilitate* (pragul minim fiind de 95%), care este dependent de valoarea lui z.

Un eșantion este cu atât mai reprezentativ, cu cât eroarea pe care o facem este mai mică, iar nivelul de încredere de încredere mai mare.

Factorii care influențează gradul de reprezentativitate

Gradul de reprezentativitate este dependent de:

eterogenitatea sau omogenitatea caracteristicilor populației din care este extras eșantionul. La același volum, calcule pe variabile omogene, produc erori de eșantionare mai mici;

mărimea eșantionului. Cu cât eșantionul este mai mare, cu atât reprezentativitatea este mai mare, dar relația nu este liniară. Creșterea numărului de indivizi din eșantion peste un anumit nivel, nu mai aduce un spor notabil de reprezentativitate. În discuția privind mărimea și reprezentativitatea eșantionului nu intervine deloc problema mărimii populației! Este fără sens să spunem: "Ce proporție din populație trebuie să aibă un eșantion ca să fie reprezentativ?" Mărimea absolută a eșantionului este cea care contează;

și *procedura de eșantionare.* Am menționat anterior faptul că putem să apreciem eroarea pe care o comitem atunci când facem estimările la nivelul populației doar în cazul în care indivizii incluși în eșantion au fost selectați corect. Numai eșantioanele aleatorii permit calculul erorilor de eșantionare.

În determinarea mărimii eșantionului aleatoriu se ține cont de resursele disponibile (financiare, umane, de timp) și de reprezentativitatea dorită. În ceea ce privește reprezentativitatea se stabilește eroarea maximă admisă și nivelul de încredere.



Exemplu

Eroarea maximă de 6 luni pentru estimarea vârstei unui eșantion a căror indivizi statistici au vârsta cuprinsă între 14 și 36 ani, adică 0.5 ani:
Pe baza celor doi parametri se calculează eroarea standard admisă. Eroarea standard admisă va fi jumătate din eroarea maximă (având în vedere pragul de încredere), adică 0.25 ani.
Se stabilește o mărime a abaterii standard a variabilei respective în populație care știm că nu va fi depășită. Abaterea standard maximă a vârstei este în situația în care plecăm de la ipoteza pesimistă privind omogenitatea, adică jumătate din indivizii statistici au 14 ani, iar jumătate au 36 de ani ($\sigma = 11\text{ani}$)
Se calculează mărimea eșantionului pe baza formulei pentru calculul erorii standard

$$n = \frac{\sigma^2}{e^2} = \frac{11^2}{0.25^2} = 1936$$

Proceduri de eșantionare

Există două mari categorii de proceduri de eșantionare: *eșantionarea aleatoare/probabilistă* și *eșantionarea nealeatoare/neprobabilistă*. Astfel vom avea două tipuri de eșantioane: eșantioane aleatorii/probabiliste și nealeatoare/neprobabiliste.

Scopul eșantionării probabiliste (denumită și aleatoare) este de a oferi cercetătorului capacitatea de a realiza inferențe precise privitoare la o populație mare pe baza unui număr mult mai mic de cazuri. George Gallup este primul care a reușit să facă predicții de mare precizie privind comportamentul electoral, prin sondaje, plecând de la teoria probabilităților.

Eșantion probabilist (aleator) este proiectat pe baza regulilor probabilității, care permite determinarea măsurii în care eșantionul reprezintă populația din care a fost selectat. Doar un eșantion aleatoriu permite calcularea erorii pe care o comitem în momentul în care estimăm valoarea la nivelul populației pornind de la valoarea obținută pe eșantion.

Pentru ca un eșantion să fie aleator, fiecare element al populației statistice trebuie să aibă o șansă egală, non-nulă și calculabilă de a fi selectat în eșantion. Selecția indivizilor care sunt incluși în eșantion trebuie să fie "la întâmplare". Probabilitatea ca un individ să fie selectat este de 1/N, unde N este mărimea populației.

Elaborarea unui eșantion aleator presupune realizarea următorilor pași:

- stabilirea *unității de analiză* și a *populației*. Populația din care se face inițial selecția trebuie să cuprindă toți indivizii relevanți pentru studiu (populație ideală), iar apoi se pot elimina anumite categorii;

- stabilirea *cadrului de eșantionare* (listă cu toți indivizii din populație. De exemplu, listele electorale, lista abonaților telefonici, evidența informatizată a populației etc.).

Calitatea unui eșantion aleator depinde de calitatea cadrului de eșantionare. În practică întâlnim mai multe limite ale cadrelor de eșantionare, dintre care amintim: alt grad de agregare decât cel dorit (indivizi în loc de gospodării), elemente în plus, elemente dublate, lipsa unor elemente (inadecvată sau incompletă). De la caz la caz, aceste probleme trebuie remediate. Cercetătorul are nevoie de ceva perspicacitate pentru a găsi cel mai potrivit cadru de eșantionare. În cazul în care nu există cadru de eșantionare, eșantionarea poate privi grupurile sau ariile în care indivizii pot fi găsiți.
- alegerea unei *proceduri de eșantionare*, astfel încât eșantionul să fie reprezentativ.

Proceduri de eșantionare aleatoare (probabiliste)

Eșantionarea simplă aleatoare reprezintă tipul ideal, fiecare individ are o șansă identică de a fi selectat în eșantion

Metoda urnei sau a tabelului cu numere aleatoare

Conform acestei proceduri fiecărui individ statistic din populație i se atribuie o bilă, aceste bile sunt introduse într-o urnă, după care se extrag rând pe rând elemente până la formarea eșantionului dorit. Dacă populația este de dimensiuni mari nu mai este necesară reintroducerea elementelor după extragere. În cazul folosirii unui tabel, pot fi generate șiruri de numere aleatoare, din care se extrag la întâmplare câteva numere.

Metoda pasului (eșantionare simplă sistematică)

Aceasta presupune existența unei liste care să conțină toți indivizii din populația vizată de cercetare. Acestor indivizi statistici li se atribuie numere de la 1 la N, după care se calculează un pas de eșantionare, care este egal cu raportul dintre mărimea populației statistice (N) și mărimea eșantionului dorit (n). Se alege apoi aleatoriu un număr cuprins între 1 și pasul de eșantionare, iar elementul selectat va fi primul din lista de eșantionare. Se adaugă apoi succesiv pasul de eșantionare până la formarea eșantionului dorit.

 Exemplu

Avem o populație de 6500 de elevi (N).
Cum extragem aleatoriu un eșantion de 1500 de elevi (n)?

$$\text{calculăm } pasul = \frac{N}{n} = \frac{6500}{1500} = 4.3$$

Avem toți elevii pe o listă cu numere de la 1 la 6500
Selectăm aleator primul număr de la care începe selecția, să zicem 5! (începem cu elevul nr. 5)
Selectăm următorul elev folosind pasul 4 … 5+4 = 9 (selectăm elevul cu numărul 9)
Selectăm următorul elev folosind pasul 4 … 9+4= 13 (selectăm elevul cu numărul 13)
Și așa mai departe, până îi selectăm pe toți cei 1500 de elevi

Eșantionarea stratificată

În cazul acestei proceduri de eșantionare se împarte populația în straturi în funcție de anumite criterii, după care se selectează aleatoriu indivizi din fiecare strat.

Stratificarea poate fi *proporțională*, când se respectă proporțiile dintre straturi din populație și *neproporțională*, când nu se respectă aceste proporții

Eșantionarea prin stratificare presupune cunoașterea distribuției populației pe straturi.

Această procedură are o reprezentativitate mai bună decât eșantionarea simplă aleatoare, deoarece cu cât populația este mai omogenă, cu atât este mai ușor să extragem eșantionul reprezentativ.

 Exemplu

Dacă avem o populație de 3200 de firme mici și mijlocii IMM. Cum extragem un eșantion reprezentativ de 620 de firme?
Știm că din cele 3200 de IMM, sunt distribuite în funcție de domeniul de activitate astfel:

	Nr. firme	% firme
industrie	570	17.81
construcții	633	19.78
comerț	1010	31.56
servicii	987	30.84
TOTAL	3200	100%

	%	Nr. firme eșantion
industrie	17.81	110
construcții	19.78	123
comerț	31.56	196
servicii	30.84	191
TOTAL EȘANTION	100%	620

Apoi folosim eșantionare aleatoare pentru a selecta indivizii din fiecare strat: vom selecta aleator 110 firme din industrie, 123 din construcții etc.

Eșantionarea cluster și eșantionarea multistadială

Se realizează când nu există cadru de eșantionare, dar indivizii sunt grupați pe categorii. Astfel, vom selecta aleator grupuri, sau mai mult, grupuri în cadrul unor grupuri. Într-o primă fază sunt selectate aleator o parte din grupurile populației vizate, după care din fiecare grup selectat în prima fază vor fi selectate tot aleator alte grupuri mai mici și așa mai departe până când se ajunge la nivelul elementului de bază din care este compusă populația vizată. De exemplu, dacă vrem să alegem un eșantion din populația unei țări, putem să selectăm o parte dintre județe, apoi localități, cartiere, străzi, blocuri până ajungem la persoane. Dacă sunt introduși în analiză toți indivizii din grup, eșantionarea se numește cluster, iar dacă sunt introduși numai o parte din indivizii statistici eșantionarea se numește multistadială.

Reprezentativitatea eșantionului construit prin această procedură este mai mică decât eșantionarea simplă aleatoare, deoarece eliminând anumite colectivități crește variația din cadrul eșantionului.

 Exemplu

Dorim să extragem un eșantion de elevi de liceu din Oradea.
Prima fază = selectăm aleator liceele care vor intra în eșantion din lista tuturor liceelor din Oradea. Faza 2 = din liceele selectate în prima fază, selectăm clasele care vor intra în eșantion. Faza 3 = din clasele selectate la faza 2, vom selecta elevii care vor fi subiecții din eșantionul nostru

Proceduri de eșantionare nealeatoare (neprobabiliste)

Eșantioanele nealeatoare sunt cele în care indivizii nu au fost aleși la întâmplare (nu a fost asigurată condiția ca fiecare să aibă șanse egale, non-nule de a fi selectat): persoane intervievate pe stradă, persoane care răspund la chestionare publicate în ziare, persoane care se oferă voluntar pentru a fi investigate etc.

Eșantionarea pe cote se folosește când nu avem un cadru de eșantionare. Această procedură este similară eșantionării stratificate: cotele indică frecvența cu care vor fi alese persoane de anumite categorii în eșantion. În interiorul straturilor indivizii nu sunt selectați aleator, selecția acestora fiind lăsată la latitudinea operatorilor de anchetă.

Eșantionarea tip bulgăre de zăpadă se folosește când nu se cunosc decât unii dintre membrii populației. De la aceștia vom afla informații pentru a identifica ulterior alte persoane din populație.

Concluzii

Eșantionarea reprezintă un set de operații cu ajutorul cărora extragem un eșantion dintr-o populație statistică.

Abaterea standard a mediilor tuturor eșantioanelor posibile selectate dintr-o populație se numește eroare standard.

Extinderea concluziilor obținute pe date la nivelul unui eșantion la nivelul populației statistice implică existența erorilor de eșantionare.

Statisticile la nivel de eșantion aproximează parametrii în zona unui interval de încredere.

Factorii care influențează gradul de reprezentativitate al eșantionului sunt: eterogenitatea sau omogenitatea caracteristicilor populației din care este extras eșantionul, mărimea eșantionului și procedura de eșantionare.

Procedurile de eșantionare se împart în 2 categorii: aleatoare și nealeatoare.

Cadrele de eșantionare reprezintă liste care conțin toți indivizii din populația statistică.

Cele mai utilizate proceduri de eșantionare aleatoare sunt: *eșantionarea simplă aleatoare, eșantionarea stratificată, eșantionarea cluster și cea multistadială*; din categoria procedurilor de eșantionare nealeatoare cele mai utilizate sunt: *eșantionarea pe cote și eșantionarea tip bulgăre de zăpadă.*

✍ *Exerciții și aplicații*

Exercițiul 1. Pe un eșantion de 900 persoane dintr-o populație s-a calculat media variabilei vârstă de 35 ani, cu abaterea standard de 12 ani. Estimați prin interval de încredere vârsta medie din populație, cu un nivel de încredere sau probabilitate de 95%.

Exercițiul 2. Pe un eșantion de 1600 studenți s-a obținut o frecvență de 20% a studenților care sunt de acord cu sistemul de acordare a burselor sociale. Estimați prin interval de încredere procentul studenților care sunt de acord cu sistemul de acordare a burselor sociale la nivelul populației statistice. Nivelul de încredere sau probabilitate este de 95%.

Exercițiul 3. Să presupunem că dorim să ne construim un eșantion utilizând "metoda pasului". Mărimea eșantionului dorit este de 400 indivizi statistici, iar cea a populaţiei statistice este de 2800. Care este mărimea pasului statistic? De la al câtelea individ statistic al populației poate începe punerea în practică a pasului de eșantionare?

TESTAREA IPOTEZELOR STATISTICE. TESTELE DE SEMNIFICAȚIE: TESTUL Z, TESTUL T, TESTUL CHI-PĂTRAT DE CONCORDANȚĂ

Obiective

Cunoașterea și înțelegerea conceptelor de: ipoteză statistică, ipoteză de nul şi ipoteză alternativă

Înțelegerea logicii testării ipotezelor

Calculul și interpretarea testelor de semnificație: testul z, testul t și testul chi-pătrat de concordanță

Aspecte introductive. Concepte

Testele de semnificație constituie elemente esențiale ale statisticii inferențiale. Ele ne ajută să răspundem la întrebări precum: Diferența dintre o valoare calculată la nivelul populației statistice și o valoare calculată la nivel de eșantion este semnificativă statistic sau ea se datorează unor fluctuații normale de eșantionare? Diferența dintre două valori provenite din două eșantioane diferite este semnificativă statistic? Eșantionul este reprezentativ pentru caracteristica dată? Diferența dintre 2 distribuții (de frecvență) este semnificativă statistic?

Testele de semnificație pot fi clasificate în două mari categorii: teste de semnificație parametrice (testează medii sau proporții) și teste de semnificație non-parametrice utilizate pentru date calitative.

Ipoteza statistică reprezintă o afirmație despre o populație care poate fi testată cu ajutorul unui eșantion aleatoriu.

Ipoteza de nul (Ho) presupune că o anumită valoare statistică calculată la nivel de eșantion (care poate fi media sau proporția) este egală cu o valoare test (a=b).

Ipoteza de nul (Ho) poate fi formulată astfel: diferența dintre cele 2 valori (sau distribuții în cazul testului chi-pătrat de concordanță) nu este semnificativă statistic.

Ipoteza alternativă (Ha) presupune că o anumită valoare statistică este diferită față de o valoare test (*nu* este egală cu această valoare; $a \neq b$).

Ipoteza alternativă (Ha) poate fi formulată astfel: diferența dintre cele două valori (sau distribuții în cazul testului chi-pătrat de concordanță) este semnificativă statistic.

Testul Z

Testul Z se aplică în cazul eșantioanelor mari, care au o distribuție normală de tip z și testează semnificația diferenței dintre două valori (a și b). Se poate vorbi despre

semnificația diferenței dintre două valori numai în cazul în care cel puțin una este rezultatul unui studiu pe bază de eșantion. Astfel rezultă două situații: o situație în care avem o valoare calculată la nivelul populației statistice (a), iar cealaltă valoare este obținută la nivel de eșantion (b), și o situație în care cele două valori (a și b) provin din două eșantioane diferite.

Testul Z exprimă diferența dintre cele două valori (a și b) în erori standard.

$$\mathbf{z} = \frac{|\boldsymbol{a} - \boldsymbol{b}|}{\boldsymbol{e}}$$

a – valoare obținută la nivelul populației statistice și b – valoare obținută la nivel de eșantion

sau

a – valoare obținută la nivel de eșantion și b – valoare obținută la nivel de eșantion

e – eroarea standard.

Testul Z ne ajută să răspundem la întrebarea dacă diferența dintre două valori (a și b) este semnificativă statistic.

Etapele de rezolvare a testului Z sunt următoarele:

1. Lansarea ipotezelor

Ho: diferența dintre cele două valori a și b nu este semnificativă statistic

Ha: diferența dintre cele două valori a și b este semnificativă statistic

2. Calcularea lui Z după formula: $\mathbf{z} = \frac{|\boldsymbol{a} - \boldsymbol{b}|}{\boldsymbol{e}}$

Pentru a calcula valoarea lui Z vom utiliza formula de calcul pentru eroarea standard:

$e = \frac{\sigma}{\sqrt{n}}$ pentru medii, respectiv $e = \sqrt{\frac{p(1-p)}{n}}$ pentru proporții.

3. Interpretarea rezultatului testului de semnificație: se compară rezultatul obținut cu o valoare critică (o valoare tabelară; tabelul standard cu valorile lui Z este prezentat în Anexa 1) corespunzătoare nivelului de probabilitate dorit.

Nivelul de probabilitate minim acceptat în științele sociale este de 95% (p=0.05). Pentru nivelul de probabilitate de 95%, valoarea critică a lui Z este de 1.96. Astfel, dacă Z>1.96, atunci Ho se respinge și Ha se acceptă, cu o probabilitate de eroare de 5%. Cu alte cuvinte, diferența dintre cele 2 valori este semnificativă statistic deoarece sunt 95% șanse ca diferența să fie reală. Dacă Z<1.96, atunci Ho se acceptă și Ha se respinge. Astfel, cu o probabilitate de eroare de 5%, diferența dintre cele 2 valori a și b nu este semnificativă statistic.

Pentru nivelul de probabilitate de 99%, valoarea critică a lui Z este de 2.58. Dacă Z>2.58, vom avea o diferență semnificativă din punct de vedere statistic între cele două valori deoarece avem o probabilitate de eroare de 1% (sunt 99% șanse ca

diferența să fie semnificativă). Dacă obținem o valoare Z<2.58, atunci diferența dintre cele două valori nu este semnificativă din punct de vedere statistic.

În situația în care cele două valori (a și b) sunt obținute la nivel de eșantion, formula pentru calcularea valorii lui Z este aceeași, dar pentru a calcula e vom avea

$e = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$, unde n1 și n2 sunt dimensiunile celor două eșantioane, iar σ_1^2 și σ_2^2pătratele abaterilor standard pentru valorile obținute la nivelul primului eșantion, respectiv cel de-al doilea eșantion.

 Exemple

Exemplul 1. La nivelul populației am obținut un procent de 28.4% persoane care s-au vaccinat cu vaccinul antigripal. În urma unui sondaj pe un eșantion de 3400 de persoane, am obținut un procent de 31.2% persoane care declară că s-au vaccinat cu serul antigripal. Dorim să verificăm dacă diferența dintre cele două valori este semnificativă din punct de vedere statistic. A crescut într-adevăr procentul populației vaccinate?
Datele problemei sunt următoarele:
valoarea la nivel de populație = 28.4%
valoarea la nivel de eșantion = 31.2%
n dimensiunea eșantionului = 3400

Pentru a testa dacă diferența dintre cele două valori este semnificativă, trebuie să calculăm valoarea lui Z.

$\mathrm{Z} = \frac{|\boldsymbol{a}-\boldsymbol{b}|}{\boldsymbol{e}}$, iar $e = \frac{\sigma}{\sqrt{n}}$

Deoarece datele sunt exprimate sub formă de procente, vom avea $\sigma = \sqrt{p(1-p)}$.

Astfel $e = \sqrt{\frac{p(1-p)}{n}}$

$$e = \sqrt{\frac{0.312(1-0.312)}{3400}} = 0.00794$$

$$Z = \frac{|28.4 - 31.2|}{0.794} = 3.52$$

Vom compara valoarea lui Z cu cea critică pentru pragul de 95%, și anume 1.96.

3.52>1.96, astfel Ho (ipoteza potrivit căreia valorile nu sunt diferite) se respinge și Ha (ipoteza potrivit căreia valorile sunt semnificativ diferite) se acceptă, cu o probabilitate de eroare de 5%. Cu alte cuvinte, diferența dintre cele 2 valori, 28.4% la nivel de populație și 31.2% la nivel de eșantion este semnificativă.

Dacă urmărim valoarea lui Z în tabelul cu distribuția normală (tabelul standard cu valorile lui Z), vom obține pentru Z=3.52 o arie între 0 și Z de 0.49978. Deoarece vorbim de o distribuție normală (simetrică la dreapta și la stânga), vom avea 0.4997*2=0.99956. Astfel, sunt 99.956% șanse ca diferența dintre cele două valori să fie diferită, un prag peste pragul minim acceptat de 95%.

Exemplul 2. Angajații din cadrul unei companii au obținut valoarea medie de 68 la un test privind competențele profesionale. La o testare realizată pe un grup de 85 angajați din cadrul companiei, valoarea medie obținută la testul privind competențele profesionale este 71, cu o abatere standard de 1.3. Dorim să verificăm dacă diferența dintre cele două valori medii obținute la test este semnificativă din punct de vedere statistic.
Datele problemei pe care le cunoaștem sunt următoarele:
valoarea la nivel de populație = 68
valoarea la nivel de eșantion = 71
abaterea standard = 1.3
n dimensiunea eșantionului =85
Vom calcula valoarea lui Z pentru a testa dacă cele două valori sunt semnificative.

$Z = \frac{|\boldsymbol{a} - \boldsymbol{b}|}{\boldsymbol{e}}$, iar $e = \frac{\sigma}{\sqrt{n}}$

$$e = \frac{\sigma}{\sqrt{n}} = \frac{1.3}{\sqrt{85}} = 0.14$$

$$Z = \frac{|\boldsymbol{a} - \boldsymbol{b}|}{\boldsymbol{e}} = \frac{|67 - 73|}{0.14} = 21.27$$

Vom compara valoarea lui Z cu cea critică pentru pragul de 95%, și anume 1.96.

21.27>1.96, astfel diferența dintre cele 2 valori, 68 la nivel de populație și 71 la nivel de eșantion este semnificativă din punct de vedere statistic.

Exemplul 3. În cadrul unui sondaj realizat pe un eșantion de 845 de elevi, am obținut un procent de 35% de elevi care declară că au participat la cel puțin o activitate extra-curriculară în anul școlar anterior. Pe un alt eșantion de 1200 de elevi, procentul elevilor care declară că au participat la cel puțin o activitate extra-curriculară este de 42%. Dorim să testăm dacă între cele două valori procentuale există o diferență semnificativă din punct de vedere statistic.

Cunoaștem următoarele valori:

n_1 = 845

procentul obținut pentru primul eșantion = 35%

n_2=1200

procentul obținut pentru al doilea eșantion = 42%

Trebuie să calculăm valoarea lui Z = $\frac{|a-b|}{e}$

Vom avea $e = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$, iar $\sigma = \sqrt{p(1-p)} = \sigma^2 = p(1-p)$

Pentru eșantionul 1 vom avea $\sigma_1^2 = 0.35(1-0.35) = 0.227$

Pentru eșantionul 2 vom avea $\sigma_2^2 = 0.42(1-0.42) = 0.243$

$$e = \sqrt{\frac{0.227}{845} + \frac{0.243}{1200}} = 0.0217$$

$$Z = \frac{|35 - 42|}{2.17} = 3.22$$

3.22>1.96. Vom respinge Ho, ipoteza potrivit căreia valorile nu sunt diferite) și vom accepta ipoteza Ha potrivit căreia valorile sunt semnificativ diferite. Din punct de vedere statistic, cele două valori obținute la nivelul celor două eșantioane sunt diferite.

Exemplul 4. Pe un eșantion de elevi am înregistrat valorile medii obținute la matematică. Pe sub-eșantionul de fete (1200 de eleve) am obținut valoarea medie de 8.85 pentru nota obținută la matematică, cu o abatere standard de 2.4. Pe sub-eșantionul format din 1500 de băieți, valoarea notei la matematică este 9.12, cu o abatere standard de 1.7. Dorim să aflăm dacă între cele două medii la matematică pe cele două sub-eșantioane există o diferență semnificativă din punct de vedere statistic. Cu alte cuvinte, vrem să verificăm dacă băieții din sub-eșantionul al doilea au note semnificativ mai bune la matematică decât cei din primul eșantion, și anume decât fetele.

Cunoaștem următoarele date ale problemei:

n_1 = 1200

valoarea medie pentru primul eșantion = 8.85

abaterea standard pentru primul eșantion = 2.4

n_2=1500

valoarea medie pentru primul eșantion = 9.12

abaterea standard pentru al doilea eșantion = 1.7

Vom calcula lui Z = $\frac{|a-b|}{e}$, unde $e = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

$$e = \sqrt{\frac{2.4^2}{1200} + \frac{1.7^2}{1500}} = 0.082.$$

$$Z = \frac{|8.85 - 9.12|}{0.082} = 3.29$$

Deoarece 3.29>1.96, vom concluziona că între cele două valori obținute pe cele două sub-eșantioane diferența este semnificativă din punct de vedere statistic. Pe datele obținute în sondaj, putem afirma că băieții au note la matematică semnificativ din punct de vedere statistic mai bune decât fetele.

Testul t (Student)

Testul t se folosește pentru eșantioanele mici (mai mici de 30 indivizi statistici), care au o distribuție a datelor de tip t (Student) și testează semnificația diferenței dintre două valori (a și b).

Formulele de calcul ale erorii standard sunt diferite, atât în cazul cu un eșantion, cât și în situația în care valorile sunt obținute pe două eșantioane.

Formula de calcul este tot $\frac{|\boldsymbol{a} - \boldsymbol{b}|}{\boldsymbol{e}}$, dar pentru a calcula eroarea standard avem:

$$e = \frac{\sqrt{\frac{\sum(x-\overline{x})^2}{n-1}}}{\sqrt{n}}$$ dacă o valoare este la nivelul populației și o valoare este la nivel de eșantion.

$$e = \sqrt{\frac{\sum(x_i-\bar{x}_1)^2+\sum(x_i-\bar{x}_2)^2}{n_1+n_2-2}}$$ dacă ambele valori sunt obținute la nivel de eșantion, unde n_1 este mărimea unui eșantion, iar n_2 este mărimea celui de-al doilea eșantion.

Pentru testarea diferenței semnificative din punct de vedere statistic, se urmărește valoarea lui t în tabelul standard (luând gradele de libertate n-1 și pragul de semnificație 0.05; tabelul cu valorile lui t este prezentat în Anexa 2). Dacă valoarea lui t calculată este mai mare decât valoarea găsită în tabelul statistic, atunci vom respinge ipoteza nulă H0 și vom accepta ipoteza alternativă Ha. În această situație diferența dintre cele două valori este semnificativă din punct de vedere statistic.

Testul chi-pătrat de concordanță

Testele z și t sunt teste parametrice și pot fi utilizate pentru date cantitative (numerice, parametrice). Când datele pe care le avem la dispoziție sunt de tip calitativ se utilizează testul chi-pătrat (χ^2)(test non-parametric).

Testul chi-pătrat are două variante: chi-pătrat de concordanță și chi-pătrat de independență (acesta este prezentat în următorul capitol).

Testul chi-pătrat de concordanță testează dacă există o diferență semnificativă între două distribuții: una observată și o distribuție teoretică.

Etapele de rezolvare a testului chi-pătrat de concordanță sunt următoarele:

1. Lansarea ipotezelor

Ho: diferența dintre cele 2 distribuții (cea observată și cea teoretică) nu este semnificativă statistic

Ha : diferența dintre cele 2 distribuții este semnificativă statistic

2. Calculul lui χ^2 după formula:

$$\chi^2 = \frac{\sum (O - T)^2}{T}$$

Frecvențele teoretice se determină pornind de la numărul total de frecvențe raportat la numărul de categorii pe care le are variabila.

3. Determinarea lui χ^2 *critic* (cu ajutorul tabelului care conține valorile critice ale lui chi-pătrat; valorile critice se regăsesc în Anexa 3). Pentru a identifica χ^2 *critic* este necesar să determinăm numărul de grade de liberate (g.l)

g.l.= n – 1, unde n reprezintă numărul de categorii a variabilei

4. Interpretarea rezultatului testului de semnificație: se compară χ^2 *calculat* cu χ^2 *critic*.

Dacă : χ^2 calculat > χ^2 critic → ipoteza de nul (Ho) se respinge, ipoteza alternativă (Ha) se acceptă (diferența dintre cele 2 distribuții este semnificativă statistic, eșantionul nu este reprezentativ pentru caracteristica dată) cu o probabilitate de eroare care reprezintă complementarul nivelului de încredere sau de probabilitate.

χ^2 calculat < χ^2 critic → ipoteza de nul (Ho) se acceptă (diferența dintre cele 2 distribuții nu este semnificativă statistic, eșantionul este reprezentativ pentru caracteristica dată) cu o probabilitate de eroare care reprezintă complementarul nivelului de încredere sau de probabilitate.



Exemple

Exemplul 1. În cadrul unui sondaj realizat pe un eșantion de 820 de studenți care au participat la cursuri online, am obținut următoarele frecvențe pentru întrebarea privind modalitatea principală de conectare la cursuri:

telefon	220
laptop	260
PC (Desktop)	140
tabletă	200

Dorim să aflăm dacă, din punct de vedere statistic, aceste răspunsuri se distribuie mai degrabă spre un anumit tip de conectare online (în sensul că studenții preferă un anumit tip mai mult decât altul) sau nu există o diferență semnificativă între diferitele modalități de conectare la cursurile online.

Pentru a calcula χ^2 *calculat*, vom determina pentru început frecvențele teoretice (=820/4=205):

	Frecvențe observat	Frecvențe teoretice
telefon	220	205
laptop	260	205
PC (Desktop)	140	205
tabletă	200	205
TOTAL	**820**	**820**

$$\chi^2 calculat = \frac{(220-205)^2}{205} + \frac{(260-205)^2}{205} + \frac{(140-205)^2}{205} + \frac{(200-205)^2}{205} = 36.59$$

Numărul de grade de libertate este 4-1=3

Vom identifica în tabelul statistic valoarea lui χ^2 *tabelar*, este 7.81.

36.59 > 7.81; χ^2 *calculat* > χ^2 *critic*

Pentru pragul de semnificație de 0.05 (cu o încredere de 95%) putem afirma că diferența dintre cele două distribuții este semnificativă din punct de vedere statistic. Vom respinge ipoteza de nul (Ho) și vom accepta ipoteza alternativă (Ha). Cu alte cuvinte, studenții care au participat la cursurile online preferă anumite modalități de accesare a cursurilor comparativ cu celelalte menționate în listă.

Exemplul 2. Pe un eșantion aleator de 1000 de persoane s-a obținut următoarea distribuție de frecvențe pentru variabila Etnie. Mai jos sunt frecvențele relative obținute la recensământul populației:

Etnia	Frecvențe relative obținute la nivel de eșantion (frecvențe observate) %	Frecvențe relative obținute la recensământ (frecvențe teoretice) %
Română	67.2	67.4
Maghiară	26	25.9
Rromă	4.8	5.0
Germană	0.2	0.2

Slovacă	1.1	1.2
Altă etnie	0.7	0.3

Pornind de la aceste date, dorim să aflăm dacă diferența dintre cele două distribuții de frecvențe este semnificativă din punct de vedere statistic. În acest scop vom folosi testul chi-pătrat de concordanță, care permite verificarea ipotezelor în această situație.

Pentru a determina χ^2 *calculat,* vom transforma pentru început frecvențele relative în frecvențe absolute.

Etnia	Frecvențe absolute (frecvențe observate) O	Frecvențe absolute (frecvențe teoretice) T
Română	672	674
Maghiară	260	259
Rromă	48	50
Germană	2	2
Slovacă	11	12
Altă etnie	7	3

$$\chi^2 calculat = \frac{(672-674)^2}{674} + \frac{(260-259)^2}{50} + \frac{(11-12)^2}{12} + \frac{(7-3)^2}{3} = 5.501$$

Numărul de grade de libertate este 6-1=5

Vom identifica în tabelul statistic valoarea lui χ^2 critic, care este 11.07

5.501 < 11.07; χ^2 *calculat* < χ^2 *critic*

Pentru pragul de semnificație de 0.05 (cu o încredere de 95%) putem afirma că diferența dintre cele două distribuții nu este semnificativă din punct de vedere statistic. Vom accepta ipoteza de nul (Ho) și vom respinge ipoteza alternativă (Ha) cu o probabilitate de eroare de 5%, ceea ce înseamnă că eșantionul este reprezentativ pentru variabila Etnie.

Concluzii

Pentru testarea ipotezelor statistice se folosesc teste de semnificație parametrice și non-parametrice.

Testele z și t se numesc teste parametrice, se referă la medii sau proporții (date cantitative, numerice, parametrice) și testează semnificația diferenței dintre două valori.

Când datele pe care le avem la dispoziție sunt de tip calitativ se utilizează testul chi-pătrat (test non-parametric). Testul este de două tipuri: testul chi-pătrat de concordanță și chi-pătrat de independență.

✍ Exerciții și aplicații

Exercițiul 1. La nivelul unei populații am înregistrat procentul de 67% de persoane care au călătorit în ultimul an în străinătate. Pe un eșantion de 1200 de persoane, procentul obținut în cazul celor care declară că au călătorit în străinătate în ultimul an este de 62%. Există o diferență semnificativă din punct de vedere statistic între cele două valori procentuale?

Exercițiul 2. În cadrul unui sondaj realizat pe un eșantion de 1500 de persoane, am obținut o valoare medie de 17.4 cu privire la numărul tranzacțiilor achitate cu cardul în ultima lună de către respondenți, cu o abatere standard de 2.5. La nivelul populației, numărul tranzițiilor cu cardul pe lună este de 15.8. Verificați dacă între cele două valori există o diferență semnificativă din punct de vedere statistic.

Exercițiul 3. La nivelul unui eșantion de 890 de subiecți am obținut valoare venitului mediu pe gospodărie de 4123 de lei, cu o abatere standard de 11.4. În cadrul unui eșantion de 640 de persoane, valoare venitului pe gospodărie este de 4566 lei, cu abaterea standard de 14.5. Testați dacă între cele două valori obținute pe cele două eșantioane există o diferență semnificativă din punct de vedere statistic.

Exercițiul 4. În cadrul unui sondaj realizat în rândul angajaților din domeniul public, am înregistrat procentul respondenților care declară că s-au gândit în ultimul an să demisioneze. Valoarea procentuală obținută în cadrul sub-eșantionului femeilor este de 13.2% (numărul de femei este de 245), iar procentul obținut în cazul bărbaților este de 11.8% (eșantionul bărbaților este de 326 de subiecți). Verificați dacă între cele două valori obținute pe cele două sub-eșantioane există o diferență semnificativă din punct de vedere statistic.

Exercițiul 5. Într-un eșantion aleator de 978 elevi de liceu s-a obținut următoarea distribuție pe naționalități. Alături sunt procentele teoretice.

Naționalitate	Frecvențe observate	Procente teoretice
Român	801	70
Maghiar	168	25
Rrom	2	1
Alta	7	4

Este semnificativă din punct de vedere statistic diferența dintre cele 2 distribuții? Argumentați folosind tabelul cu valorile critice ale lui chi-pătrat.

Exercițiul 6. Într-un eșantion aleator de 500 persoane s-a obținut următoarea distribuție pe religii. Alături sunt procentele teoretice. Este eșantionul reprezentativ pentru caracteristica dată? Argumentați folosind tabelul cu valorile critice ale lui chi-pătrat.

Religia	Procente obținute la nivel de eșantion	Procente teoretice
Ortodoxă	60	58
Catolică	10	12
Baptistă	6	5
Penticostală	4	6
Alta	20	19

Exercițiul 7. În cadrul unui sondaj am obținut următoarea distribuție pentru răspunsurile la întrebarea *Unde iei masa de prânz de obicei?*

La cantina universității	42
În oraș	36
Acasă	52
În altă locație	30

Este una dintre aceste locații preferată de către studenții respondenți în cadrul sondajului sau răspunsurile lor se distribuie omogen? Este o diferență statistică semnificativă între distribuția răspunsurilor studenților și o distribuție teoretică?

ASOCIEREA VARIABILELOR CALITATIVE. TESTUL CHI-PĂTRAT DE INDEPENDENȚĂ

Obiective

Prezentarea unui tabel de contingență

Testarea relației dintre două variabile calitative – calculul și interpretarea lui chi-pătrat de independență

Prezentarea coeficienților care exprimă intensitatea relației dintre variabilele calitative

Ce este tabelul de contingență

Relația dintre două variabile calitative este prezentată într-un tabel de contingență, tabel cu dublă intrare, în care valorile uneia dintre variabile apar pe linii, iar valorile celeilalte apar pe coloane.

Tabelul de contingență este un tabel de asociere care oferă informații despre 2 tipuri de distribuții de frecvențe: marginale (distribuții de frecvențe ale variabilelor) și interioare.

Tabel de contingență		**VARIABILA A**					
		Categoria 1 var. A	Categoria 2 var. A	Categoria 3 var. A	…	Categoria n var. A	**TOTAL**
VARIABILA B	Categoria 1 var. B	O_{A1B1}	O_{A2B1}	O_{A3B1}	…	O_{AnB1}	$\mathbf{N_{B1.}}$
	Categoria 2 var. B	O_{A1B2}	O_{A2B2}	O_{A3B2}	…	O_{AnB2}	$\mathbf{N_{B2.}}$
	Categoria 3 var. B	O_{A1B3}	O_{A2B3}	O_{A3B2}	…	O_{AnB3}	$\mathbf{N_{B3.}}$
	…	…	…	…	…	…	
	Categoria m var. B	O_{A1Bm}	O_{A2Bm}	O_{A3Bm}	…	O_{AnBm}	$\mathbf{N_{Bm.}}$
	TOTAL	$\mathbf{N_{A1.}}$	$\mathbf{N_{A2.}}$	$\mathbf{N_{A3.}}$		$\mathbf{N_{An.}}$	**N**

n – numărul de categorii ale variabilei A

m – numărul de categorii ale variabilei B

O_{A1B1}, O_{A2B1}, O_{A3B1}, O_{A1B2}, O_{A2B2} etc. – frecvențe interioare (distribuția în funcție de variabilele A şi B)

$N_{A1.}$, $N_{A2.}$, $N_{A3.}$, $N_{An.}$ – frecvențele marginale pentru variabila A

$N_{B1.}$, $N_{B2.}$, $N_{B3.}$, $N_{Bm.}$ – frecvențele marginale pentru variabila B

N – numărul total de indivizi statistici

 Exemplu

Pe un eșantion de 550 de subiecți (studenți), am înregistrat răspunsurile la întrebarea *Cât de frecvent ai folosit biblioteca universității pentru studii și lectură?* Distribuția răspunsurilor în funcție de nivelul de studiu se prezintă astfel:

		Variabila NIVEL DE STUDIU		
		LICENTA	MASTER	**TOTAL**
Variabila BIBLIOTECA: Cât de frecvent ai folosit biblioteca universității?	niciodată	144	58	**202**
	mai rar	61	59	**120**
	o dată pe lună	39	14	**53**
	de câteva ori pe lună	27	15	**42**
	o dată pe săptămână	23	18	**41**
	de mai multe ori pe săptămână	27	21	**48**
	zilnic	28	16	**44**
	TOTAL	**349**	**201**	**550**

Pe ultima coloană, respectiv pe ultimul rând sunt frecvențele marginale. Ultima coloană cuprinde numărul total de cazuri pe fiecare categorie a variabilei BIBLIOTECA: 202 studenți au răspuns "niciodată", 120 au răspuns "mai rar" etc.. Ultimul rând cuprinde numărul total de cazuri pe fiecare NIVEL DE STUDIU: 349 de studenți sunt la nivel de licență, 201 sunt la nivel de masterat.
Celelalte valori din tabel sunt frecvențele interioare. Vom citi astfel: 144 de studenți de la nivel de licență răspund la întrebarea referitoare la BIBLIOTECĂ "niciodată", 61 studenți de la nivel de licență răspund "mai rar", 39 de studenți de la nivel de licență răspund "o dată pe lună" etc. 58 de studenți de la nivel de masterat răspund la întrebarea referitoare la BIBLIOTECĂ "niciodată", 59 studenți de la nivel de masterat răspund "mai rar", 14 de studenți de la nivel de masterat răspund "o dată pe lună" etc..

Testul chi-pătrat de independență

Testul este utilizat pentru a verifica dacă două variabile calitative sunt sau nu sunt asociate. Testul chi-pătrat de independență își propune să răspundă la întrebarea: există asociere între două variabile calitative?

Etapele de rezolvare ale testului χ^2 de independență sunt următoarele:

1. Lansarea ipotezelor

Ho: nu există asociere între două variabile calitative.

Ha: există asociere între două variabile calitative.

2. Calculul lui chi-pătrat după formula:

$$\chi^2 = \frac{\sum (O - T)^2}{T}$$

O - frecvențe observate

T - frecvențe teoretice/așteptate

În determinarea frecvențelor teoretice în ipoteza independenței plecăm de la probabilități: două evenimente A și B sunt independente dacă P(A și B) = P(A) x P(B).

Pornind de la tabelul de contingență (care cuprinde frecvențele observate O), putem genera un tabel teoretic (tabel de independență) care conține frecvențele teoretice (T).

Tabel de independență		**VARIABILA A**					
		Categoria 1 var. A	Categoria 2 var. A	Categoria 3 var. A	…	Categoria n var. A	**TOTAL**
VARIABILA B	Categoria 1 var. B	T_{A1B1}	T_{A2B1}	T_{A3B1}	…	T_{AnB1}	$\mathbf{N_{B1.}}$
	Categoria 2 var. B	T_{A1B2}	T_{A2B2}	T_{A3B2}	…	T_{AnB2}	$\mathbf{N_{B2.}}$
	Categoria 3 var. B	T_{A1B3}	T_{A2B3}	T_{A3B2}	…	T_{AnB3}	$\mathbf{N_{B3.}}$
	…	…	…	…	…	…	
	Categoria m var. B	T_{A1Bm}	T_{A2Bm}	T_{A3Bm}	…	T_{AnBm}	$\mathbf{N_{Bm.}}$
	TOTAL	$\mathbf{N_{A1.}}$	$\mathbf{N_{A2.}}$	$\mathbf{N_{A3.}}$		$\mathbf{N_{An.}}$	**N**

Frecvențe teoretice din acest tabel sunt calculate pornind de la frecvențele marginale după următoarea formulă:

$$T = \frac{frecvența\ marginală\ pe\ coloană * frecvența\ marginală\ pe\ linie}{total}$$

Astfel:

$$T_{A1B1} = \frac{N_{A1}*N_{B1}}{N}\ ;\ T_{A1B2} = \frac{N_{A1}*N_{B2}}{N}\ ;\ T_{A1B3} = \frac{N_{A1}*N_{B3}}{N}\ ;\ T_{A2B1} = \frac{N_{A2}*N_{B1}}{N}\ \text{etc.}$$

Pe baza valorilor din cele două tabele (tabelul de contingență și tabelul teoretic/de independență), vom calcula χ^2 *calculat* după formula de mai sus

$$\chi^2 = \frac{\sum (O - T)^2}{T}$$

Astfel:

$$\chi^2 = \frac{(O_{A1B1} - T_{A1B1})^2}{T_{A1B1}} + \frac{(O_{A1B2} - T_{A1B2})^2}{T_{A1B2}} + \frac{(O_{A1B3} - T_{A1B3})^2}{T_{A1B3}} + \cdots + \frac{(O_{A1Bm} - T_{A1Bm})^2}{T_{A1Bm}} + \frac{(O_{A2B1} - T_{A2B1})^2}{T_{A2B1}} + \frac{(O_{A2B2} - T_{A2B2})^2}{T_{A2B2}} + \frac{(O_{A2B3} - T_{A2B3})^2}{T_{A2B3}} + \cdots + \frac{(O_{A2Bm} - T_{A2Bm})^2}{T_{A2Bm}} + \cdots + \frac{(O_{AnB1} - T_{AnB1})^2}{T_{AnB1}} + \frac{(O_{AnB2} - T_{AnB2})^2}{T_{AnB2}} + \frac{(O_{AnB3} - T_{AnB3})^2}{T_{AnB3}} + \cdots + \frac{(O_{AnBm} - T_{AnBm})^2}{T_{AnBm}}$$

3. Identificarea lui χ^2 *critic* (cu ajutorul tabelului care conține valorile critice ale lui chi-pătrat; Anexa 3). Pentru a identifica χ^2 *critic* este necesar să determinăm numărul de grade de liberate (g.l)

gl= (l– 1) * (c – 1)

unde l este numărul liniilor și c este numărul coloanelor tabelului de contingență

Astfel, gradele de libertate ale tabelului care cuprind frecvențele variabilelor A și B din tabelul anterior: gl= (m– 1) * (n – 1)

4. Interpretarea rezultatului testului: se compară χ^2 *calculat* cu χ^2 *critic*

Dacă**:** χ^2 *calculat* > χ^2 *critic* → ipoteza de nul (Ho) se respinge, ipoteza alternativă (Ha) se acceptă (există asociere între cele 2 variabile calitative) cu o probabilitate de eroare care reprezintă complementarul nivelului de încredere sau de probabilitate.

Dacă**:** χ^2 *calculat* < χ^2 *critic* → ipoteza de nul (Ho) se respinge, ipoteza alternativă (Ha) se acceptă (nu există asociere între cele 2 variabile calitative) cu o probabilitate de eroare care reprezintă complementarul nivelului de încredere sau de probabilitate.

 Exemplu

Pornind de la datele obținute în cadrul sondajului cu cei 550 de studenți, dorim să aflăm dacă există o asociere între cele două variabile: frecventarea bibliotecii și nivelul de studiu.
Tabelul care cuprinde frecvențele observate este cel prezentat în exemplul anterior:

Tabelul de contingență (frecvențele observate)		**Variabila NIVEL DE STUDIU**		
		LICENTA	MASTER	**TOTAL**
Variabila BIBLIOTECA: Cât de frecvent ai folosit biblioteca universității?	niciodată	144	58	**202**
	mai rar	61	59	**120**
	o dată pe lună	39	14	**53**
	de câteva ori pe lună	27	15	**42**
	o dată pe săptămână	23	18	**41**
	de mai multe ori pe săptămână	27	21	**48**
	zilnic	28	16	**44**
	TOTAL	**349**	**201**	**550**

Ipotezele sunt următoarele:

Ho: nu există asociere între două variabile calitative.

Ha: există asociere între două variabile calitative.

Pornind de la frecvențele marginale (pe linii și coloane) din tabelul de contingență, vom calcula frecvențele teoretice:

Tabelul de independență (frecvențele teoretice)		**Variabila NIVEL DE STUDIU**		
		LICENTA	MASTER	**TOTAL**
Variabila BIBLIOTECA: Cât de frecvent ai folosit biblioteca universității?	niciodată	128.18	73.82	**202**
	mai rar	76.15	43.85	**120**
	o dată pe lună	33.63	19.37	**53**
	de câteva ori pe lună	26.65	15.35	**42**
	o dată pe săptămână	26.02	14.98	**41**
	de mai multe ori pe săptămână	30.46	17.54	**48**
	zilnic	27.92	16.08	**44**
	TOTAL	**349**	**201**	**550**

$$\mathbf{128.18} = \frac{349 * 202}{550}; \mathbf{76.15} = \frac{349 * 120}{550}; \mathbf{33.63} = \frac{349 * 53}{550};$$

$$\mathbf{26.65} = \frac{349 * 42}{550}; \ldots \mathbf{73.82} = \frac{201 * 202}{550}; \mathbf{43.85} = \frac{201 * 120}{550} etc.$$

Calculăm χ^2 *calculat* după formula:

$$\chi^2 = \frac{\sum (O - T)^2}{T}$$

$$\chi^2 = \frac{(144-128.18)^2}{128.18} + \frac{(61-76.15)^2}{76.15} + \frac{(39-33.63)^2}{33.63} + \frac{(27-26.65)^2}{26.65} + \frac{(23-26.02)^2}{26.02} + \frac{(27-30.46)^2}{30.46} + \frac{(28-27.92)^2}{27.92} + \frac{(58-73.82)^2}{73.82} + \frac{(59-43.85)^2}{43.85} + \frac{(14-19.37)^2}{19.37} + \frac{(15-15.35)^2}{15.35} + \frac{(18-14.98)^2}{14.98} + \frac{(21-17.54)^2}{17.54} + \frac{(16-16.08)^2}{16.08} = 17.98$$

χ^2 *calculat* = 17.98

Calculăm numărul gradelor de libertate:

g.l.=(număr de linii – 1)*(număr de coloane – 1) = (7-1)*(2-1)=6

Identificăm χ^2 *critic* (în tabelul care conține valorile critice ale lui chi-pătrat) pentru pragul de semnificație de p=0.05 și g.l. = 6

χ^2 *critic* = 12.59

17.98 > 12.59. Ca atare, χ^2 *calculat* > χ^2 *critic*

Potrivit rezultatelor testului, vom respinge ipoteza de nul (Ho) și vom accepta ipoteza alternativă (Ha). Concluzia este că există asociere între cele 2 variabile calitative: frecventarea bibliotecii de către studenți și nivelul de studiu.

Coeficienți care exprimă intensitatea relației dintre variabilele calitative

Testul chi-pătrat de independență ne oferă informații privind existența sau inexistența unei relații de asociere între două variabile calitative, dar nu și cât de intensă este această asociere atunci când ea există. Pentru a determina care este intensitatea relației dintre două variabile calitative se apelează la coeficienți care sunt diferiți pentru cele două tipuri de variabile calitative nominale, respectiv ordinale.

Coeficienți pe baza lui χ^2

Utilizând valoarea lui χ^2 nu putem stabili care dintre variabile este cea dependentă și care este cea independentă. În plus, valoarea lui χ^2 nu ne indică informații cu privire la intensitatea relației. Mărimea lui χ^2 nu exprimă intensitatea deoarece valoarea lui

χ^2 depinde de dimensiunea tabelului de contingență (de numărul de grade de libertate), respectiv de dimensiunea eșantionului.

În acest sens, pentru a avea o imagine cu privire la intensitatea relației de asociere, sunt utilizați coeficienți calculați pe baza lui χ^2.

Coeficientul φ (PHI)

$$\varphi = \sqrt{\frac{\chi^2}{n}}$$

Coeficientul V al lui Cramer

$V = \sqrt{\frac{\chi^2}{n(s-1)}}$ unde s este cel mai mic din numărul liniilor și coloanelor

Coeficientul de contingență al lui Pearson

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}}$$

Acești indicatori au valori încadrate în intervalul (0,1).

 Exemplu

Folosind rezultatele din exercițiul anterior cu privire la relația de asociere dintre frecventarea bibliotecii și nivelul de studiu, vom calcula valorile coeficienților pe baza lui χ^2.

χ^2 *calculat* = 17.98

$$\varphi = \sqrt{\frac{\chi^2}{n}} = \sqrt{\frac{17.98}{550}} = 0.18$$

$$V = \sqrt{\frac{\chi^2}{n(s-1)}} = \sqrt{\frac{17.98}{550(2-1)}} = 0.18$$

$$C = \sqrt{\frac{\chi^2}{n+\chi^2}} = \sqrt{\frac{17.98}{550+17.98}} = 0.18$$

Potrivit acestor valori, observăm că relația de asociere este destul de slabă.

Coeficientul lambda (λ) pentru variabile nominale

λ este utilizat pentru a determina intensitatea relației dintre două variabile nominale și se calculează după formula de mai jos:

$$\lambda = \frac{e1 - e2}{e1}$$

Semnificația coeficientului - reducerea proporțională a erorii în predicția valorilor variabilei efect (dependente) prin cunoașterea valorilor variabilei cauză (independente). λ variază între 0 și 1.

Unul dintre principalele dezavantaje ale acestui coeficient este că, dacă una dintre valorile variabilei dependente este mult mai frecventă decât celelalte, coeficientul ia valoarea 0 chiar dacă există asociere.

Coeficientul τ al lui Kendall pentru variabile ordinale

$\tau = \frac{nc - nd}{nt}$, unde nc reprezintă numărul perechilor concordante, nd – numărul perechilor discordante, iar nt – numărul total de perechi

Valorile lor pot fi ordonate. Se poate vorbi de semn al asocierii, de direcție, coeficienții de acest fel variind între -1 și +1.

Asociere pozitivă între variabilele ordinale X și Y: valorilor mari ale variabilei X le corespund valori mari ale variabilei Y iar celor mici, în mod analog, le corespund valori mici.

Asociere negativă: valorilor mici ale variabilei X le corespund valori mari ale lui Y iar celor mari le corespund valori mici ale lui Y.

Variabilele trebuie să fie ambele ordinale, iar valorile ordonate în același fel.

Concluzii

Tabelul de contingență (asociere) este un tabel cu dublă intrare, în care valorile uneia dintre variabile apar pe linii, iar valorile celeilalte variabile apar pe coloane.

Testarea asocierii dintre două variabile calitative se realizează cu ajutorul testului chi-pătrat de independență.

Coeficienții utilizați pentru a determina intensitatea relației dintre variabilele calitative sunt: coeficientul φ (PHI), coeficientul V al lui Cramer, coeficientul de contingență al lui Pearson, coeficientul λ pentru variabilele nominale și coeficientul τ al lui Kendall pentru variabilele ordinale.

✍ Exerciții și aplicații

Exercițiul 1. Într-un eșantion aleator de 978 de elevi de liceu, relația dintre gen și tipul liceului este prezentată în tabelul de contingență de mai jos:

Gen / Tipul liceului	Liceu teoretic sau Colegiu Național	Grup școlar
Masculin	179	256
Feminin	338	205

Există asociere între cele 2 variabile calitative (gen și tipul liceului)?

Argumentați folosind tabelul cu valorile critice ale lui chi-pătrat.

Exercițiul 2. Testați dacă există o asociere între cele două variabile (timpul alocat pregătirii lecțiilor și gen), pornind de la valorile observate din tabelul de mai jos:

		Gen		Total
		masculin	feminin	
Într-o zi obișnuită a săptămânii, câte ore aloci pentru pregătirea cursurilor, seminariilor, laboratoarelor?	Peste 3 ore	137	379	516
	2-3 ore	186	422	608
	1-2 ore	211	370	581
	0,5 -1 ora	132	197	329
	Sub o jumătate de ora	95	107	202
Total		761	1475	2236

Exercițiul 3. În cadrul unui sondaj am obținut răspunsuri la întrebarea *În ce măsură sunteți preocupat de alimentația dvs.?* Distribuția răspunsurilor în funcție de mediul de rezidență este prezentată în următorul tabel:

	URBAN	RURAL	**TOTAL**
În foarte mare măsură	126	111	237
În mare măsură	125	75	200
În mică măsură	85	29	114
În foarte mică măsură	55	32	87
TOTAL	391	247	638

Există o asociere între cele două variabile?

TESTE GRILĂ ȘI APLICAȚII DE SINTEZĂ

Teste grilă propuse

1. Distribuția variabilei ”stare civilă” într-un grup de 7 indivizi statistici este: necăsătorit, necăsătorit, necăsătorit, căsătorit, căsătorit, văduv, divorțat. Cum poate fi descrisă tendința centrală pentru această variabilă:

a. Mod=3
b. Mod=necăsătorit
c. Mod =necăsătorit, mediana=căsătorit
d. Mod=3, mediana=2

2. Operațiile de adunare/scădere sunt permise pentru variabilele:

a. calitative
b. ordinale
c. de interval
d. de raport

3. Care dintre următoarele propoziții nu sunt proprietăți ale mediei aritmetice:

a. Media este o valoare pe care nu o ia nici un individ statistic
b. Media este cuprinsă între valoarea minimă și maximă a seriei de date
c. Media se exprimă în aceeași unitate de măsură ca și valorile variabilei

4. Care dintre indicatorii tendinței centrale pot fi determinați pentru variabilele cantitative:

a. doar media
b. doar modul și mediana
c. modul, mediana, media
d. modul, mediana, media, abaterea standard

5. Frecvența relativă poate fi determinată pentru variabilele:

a. nominale
b. ordinale
c. de interval
d. de raport

6. Media aritmetică nu poate fi determinată pentru variabilele:

a. Calitative
b. Cantitative
c. De interval
d. De raport

7. Valoarea unei variabile calculată la nivelul populației statistice se numește:

a. Unitate statistică

b. Frecvență relativă
c. Parametru
d. Statistică la nivel de eșantion

8. Intervalul modal este:
a. Cel corespunzător primei frecvențe cumulate care depășește jumătate din observații
b. Cel care are frecvența cea mai mare
c. Cel care împarte seria ordonată de date în 2 părți egale

9. Distribuția variabilei ”etnie” într-un grup de 7 indivizi statistici este: român, maghiar, rrom, maghiar, român, român, altă etnie. Cum poate fi descrisă tendința centrală pentru această variabilă:
a. Mod=3
b. Mod=român
c. Mod =român, mediana=maghiar
d. Mod=3, mediana=2

10. Pentru a împărți o serie de date în patru părți egale avem nevoie de :
a. patru valori quartile
b. trei valori quartile
c. quartila 1 și quartila 2

11. Ordonarea valorilor sau categoriilor variabilei nu este permisă pentru variabilele:
a. nominale
b. ordinale
c. de interval
d. de raport

12. Care dintre indicatorii tendinței centrale pot fi determinați pentru variabilele cantitative:
a. doar media
b. doar modul și mediana
c. modul, mediana, media
d. modul, mediana, media, abaterea standard

13. Abaterea standard poate fi determinată pentru variabilele:
a. nominale
b. ordinale
c. de interval
d. de raport

14. Media aritmetică nu poate fi determinată pentru variabilele:
- a. Nominale
- b. Ordinale
- c. De interval
- d. De raport

15. Valoarea unei variabile calculată la nivelul de eșantion se numește:
- a. Unitate statistică
- b. Frecvență relativă
- c. Parametru
- d. Statistică la nivel de eșantion

16. Variabilelor cantitative le putem aplica:
- a. Statistica parametrică și non-parametrică
- b. Doar statistica parametrică
- c. Doar statistica non-parametrică

17. În cadrul seriei de date: 1, 2, 2, 3, 3, 2 mediana este:
- a. 2.5
- b. 2
- c. 3
- d. 2.3

18. Am obținut următoarele răspunsuri la întrebarea ”De câte ori ați prestat ore suplimentare în ultima lună” : 4, 6, 1, 5, 1, 6, 1, 4, 5, 2, 6, 2, 5, 3, 3, 3, 5, 4, 6, 4, 7, 4. Valorile quartile sunt:
- a. 3, 4, 5
- b. 4, 5, 6
- c. 5, 6, 7
- d. 25%, 50%, 75%

19. Se dau următoarele valori: 5, 4, 6, 4, 5, 6, 7, 6, 4, 6, 8, 6, reprezentând salariile brute orare obținute de către un grup de angajați ai unei organizații private. Frecvența relativă a valorii 4 este:
- a. 0.16
- b. 0.41
- c. 0.25
- d. 0.33

20. Se dau următoarele valori: 8, 10, 9, 7, 8, 10, 6, 6 . Valoarea abaterii standard este:
- a. 4.24
- b. 1.8
- c. 1.5

d. 2.25

21. Media aritmetică a valorilor: 10, 6, 7, 8, 7, 5, 9, 10 este:

a. 7.75
b. 8.75
c. 7
d. 8

22. Care dintre următoarele propoziții nu sunt adevărate pentru o distribuție normală:

a. distribuția este simetrică în jurul mediei
b. media este egală cu 0 și abaterea standard este egală cu 1
c. media este egală cu mediana
d. distribuția este unimodală

23. Eroarea standard semnifică:

a. eroarea unui eșantion aleator
b. abaterea standard a distribuţiei de eşantionare a mediei
c. greşeala care se face de obicei la un eşantion aleator
d. media abaterilor standard ale distribuţiei de eşantionare a mediei

24. Se dau următoarele valori ale unei variabile cantitative: 4, 5, 6, 6, 7, 8, 7, 5. Care este abaterea standard a acestei variabile?

a. 1.22
b. 1.5
c. 2.22
d. 1.75

25. Pe un eșantion de 1000 subiecți s-a calculat media variabilei vârstă de 45 ani și abaterea standard de 17 ani. Care este intervalul de încredere al estimării acestui parametru la pragul de încredere de 95%.

a. 45 +/- 1.96
b. 45 +/- 1.052
c. 45 +/- 0.537
d. 45 +/- 17

26. Se dau următoarele valori: 4, 6, 5, 4, 6, 7, 5, 8, 4, 6, 7, 5, 6, 8 reprezentând salariile brute orare (exprimate în u.m.) ale angajaților unei organizații. Care este ponderea (proporția) angajaților care au salariul brut orar de 6 u.m. sau mai puțin?

a. 28.57%
b. 71.42%
c. 42.85%
d. 85.71%

27. Se dau următoarele valori: 1, 2, 2, 3, 4, 4, 4, 2. Mediana distribuției seriei de date este:

a. 2.5
b. 3.5
c. 2
d. 3 și 4

28. Din 800 cadre didactice, 432 au spus că au intenția de a susține candidatul A la alegerile pentru funcția de rector a unei universitășii. Eroarea de eșantionare pentru un nivel de încredere de 95% este:

a. 5%
b. 3.3%
c. 0.54%
d. 4.4%

29. Între un eşantion simplu aleator şi unul cluster sau multistadial, este mai reprezentativ:

a. Cel simplu aleator
b. Eșantionul cluster sau multistadial
c. Reprezentativitatea nu poate fi determinată
d. Au același grad de reprezentativitate

30. Care dintre următoarele caracteristici califică o procedură de eşantionare ca fiind probabilistă:

a. toate categoriile au o reprezentare egală în eșantion
b. toţi indivizii din populaţie au aceeaşi probabilitate de a fi selectaţi în eşantion
c. probabilitatea fiecărui individ din populaţie de a fi selectat este calculabilă şi diferită de zero
d. alegerea indivizilor în eşantion se face fără prejudecăţi

31. Un grup de studenți au luat următoarele note la statistică: 7,9,10,5,8,6,9,7,8,6. Care este abaterea standard a celor 10 note?

a. 1.5
b. 2.25
c. 2.5
d. 1.6

32. Pe un eșantion de 1600 studenți s-a calculat media variabilei note de 7,80 și abaterea standard de 1,6. Care este intervalul de încredere al estimării acestui parametru la pragul de încredere de 95%.

a. 7.8 +/- 1.96
b. 7.8 +/- 0.078

c. 7.8 +/- 0.040
d. 7.8 +/- 1.6

33. Se dau următoarele valori: 4, 5, 4, 6, 7, 5, 8, 4, 6, 7, 5, 6 reprezentând salariile brute orare (exprimate în u.m.) ale angajaților unei organizații. Care este ponderea (proporția) angajaților care au salariul brut orar de 6 u.m. sau mai puțin?
a. 25%
b. 75%
c. 50%
d. 91.66%

34. Se dau următoarele valori: 1, 2, 2, 3, 4, 4, 5, 1. Mediana distribuției seriei de date este:
a. 2.5
b. 3.5
c. 4.5
d. 3 și 4

35. Din 800 cadre didactice, 432 au spus că au intenția de a susține candidatul A la alegerile pentru funcția de rector a unei universitățiii. Eroarea de eșantionare pentru un nivel de încredere de 95% este:

a. 5%
b. 3.3%
c. 0.54%
d. 4.4%

36. Între un eşantion simplu aleator şi unul format prin stratificare, este mai reprezentativ:

a. Cel simplu aleator
b. Cel format prin stratificare
c. Reprezentativitatea nu poate fi determinată
d. Au același grad de reprezentativitate

37. Din 700 cadre didactice, 425 au spus că au intenția de a susține candidatul A la alegerile pentru funcția de rector a unei universitățiii. Eroarea de eșantionare pentru un nivel de încredere de 95% este:

a. 5%
b. 3.5%
c. 0.6%
d. 1.8 %

38. Între un eşantion pe cote şi unul tip bulgăre de zăpadă, este mai reprezentativ:

a. Eșantionul pe cote
b. Eșantionul tip bulgăre de zăpadă
c. Reprezentativitatea nu poate fi determinată
d. Au același grad de reprezentativitate

39. Distribuția variabilei ”religie” într-un grup de 7 indivizi statistici este: ”ortodox”, ”catolic”, ”reformat”, ”catolic”, ”ortodox”, ”ortodox”, ”altă religie”. Cum poate fi descrisă tendința centrală pentru această variabilă:

a. Mod=3
b. Mod=ortodox
c. Mod =ortodox, mediana=catolic
d. Mod=3, mediana=2

Aplicații de sinteză rezolvate

Problema 1. Se dă următorul tabel reprezentând distribuția unui eșantion de studenți după variabila vârstă:

Vârsta	**Număr de studenți (frecvența absolută)**
18	16
19	42
20	32
21	16
22	4
Total	110

Determinați:

A) Indicatorii tendinței centrale (Mo, Me, M)

B) Ponderea sau proporția studenților care au 21 ani

C) Ponderea sau proporția studenților care au 21 ani sau mai puțin

D) Conform unor date existe la Secretariatul facultății vârsta medie a studenților este 19.7 ani. Să se testeze semnificația diferenței dintre vârsta medie a studenților la nivel de eșantion și vârsta medie a studenților la nivelul populației statistice.

Rezolvare:
A) Indicatorii tendinței centrale sunt modul, mediana și media aritmetică.
Modul = 19 (pentru că este valoarea variabilei care are frecvența cea mai mare)
Mediana = 19 (pentru că este valoarea variabilei corespunzătoare primei frecvențe cumulate care depășește jumătate din observații)
Pentru a determina media folosim formula de mai jos:

$$\bar{x} = \frac{\sum x_i * f_i}{\sum f_i}$$

$$\bar{x} = \frac{18 * 16 + 19 * 42 + 20 * 32 + 21 * 16 + 22 * 4}{110} = 19.54\ ani$$

B) Pentru a determina ponderea sau proporția studenților care au 21 ani vom calcula frecvența relativă

$$fr_i = \frac{f_i}{\sum f_i} * 100 \ = \frac{16}{110} * 100 = 14.54\%$$

Astfel, 15.54% dintre studenți au 21 ani.

C) Pentru a determina ponderea sau proporția studenților care au 21 ani sau mai puțin va trebui să calculăm frecvențele relative cumulate

Valoarea variabilei x	**Frecvențe absolute** f_i	**Frecvențe relative (%)** fr_i	**Frecvențe cumulate**
18	16	14.55	14.55
19	42	38.18	52.73
20	32	29.09	81.82
21	16	14.55	96.37
22	4	3.63	100
	N=110	100%	

Ponderea sau proporția studenților care au 21 ani sau mai puțin este de 96.37% (14.55+38.18+29.09+14.55=96.37%)

D) Pentru a testa semnificația diferenței dintre vârsta medie a studenților la nivel de eșantion și vârsta medie a studenților la nivelul populației statistice vom folosi testul z.

Datele problemei pe care le cunoaștem sunt următoarele:

valoarea la nivel de populație = 19.7

valoarea la nivel de eșantion = 19.54

n dimensiunea eșantionului =110 studenți

Vom calcula valoarea lui Z pentru a testa dacă cele două valori sunt semnificative.

$$Z = \frac{|\boldsymbol{a} - \boldsymbol{b}|}{\boldsymbol{e}}, \text{ iar } e = \frac{s}{\sqrt{n}}$$

$$s = \sqrt{s^2} = \sqrt{\frac{\sum (x_i - \bar{x})^2 * f_i}{N-1}}$$

$$s^2 = \frac{(18-19.54)^2*16+(19-19.54)^2*42+(20-19.54)^2*32+(21-19.54)^2*16+(22-19.54)^2*4}{110-1}$$

$$s^2 = \frac{115.27}{109} = 1.05$$

Abaterea standard este radical din varianță:

$$s = \sqrt{s^2} = \sqrt{1.05} = 1.02$$

$$e = \frac{s}{\sqrt{n}} = \frac{1.02}{\sqrt{110}} = 0.0973$$

$$Z = \frac{|\boldsymbol{a} - \boldsymbol{b}|}{\boldsymbol{e}} = \frac{|19.7 - 19.54|}{0.0973} = 1.64$$

Vom compara valoarea lui Z cu cea critică pentru pragul de 95%, și anume 1.96.

1.64 < 1.96, astfel diferența dintre cele 2 valori, 19.7 la nivel de populație și 19.54 la nivel de eșantion nu este semnificativă din punct de vedere statistic.

Problema 2. Pe un eşantion aleator format din 200 de gospodării s-au obţinut următoarele date:

Număr de copii din gospodărie	**Număr de gospodării (frecvența absolută)**
0	20
1	50
2	100
3	15
4	10
5	5
Total	200

Cerințe**:**

A) Determinați indicatorii tendinței centrale

B) Determinați ponderea (proporția) gospodăriilor care au 3 copii

C) Determinați ponderea (proporția) gospodăriilor care au 3 copii sau mai puțin

D) Calculați abaterea standard

E) Estimaţi prin interval de încredere numărul mediu de copii din gospodărie la nivelul populaţiei statistice; nivelul de încredere este de 95%.

Rezolvare:

A) Indicatorii tendinței centrale sunt modul, mediana și media aritmetică.

Modul = 2 (pentru că este valoarea variabilei care are frecvența cea mai mare)
Mediana = 2 (pentru că este valoarea variabilei corespunzătoare primei frecvențe cumulate care depășește jumătate din observații)
Pentru a determina media folosim formula de mai jos:

$$\bar{x} = \frac{\sum x_i * f_i}{\sum f_i}$$

$$\bar{x} = \frac{0 * 20 + 1 * 50 + 2 * 100 + 3 * 15 + 4 * 10 + 5 * 5}{200} = 1.8$$

B) Pentru a determina ponderea (proporția) gospodăriilor care au 3 copii vom calcula frecvența relativă

$$fr_i = \frac{f_i}{\sum f_i} * 100 \ = \frac{15}{200} * 100 = 7.5\%$$

Astfel, 7.5% dintre gospodării au 3 copii.

C) Pentru a determina ponderea sau proporția gospodăriilor care au 3 copii sau mai puțin va trebui să calculăm frecvențele relative cumulate

Valoarea variabilei x	**Frecvențe absolute f_i**	**Frecvențe relative (%) fr_i**	**Frecvențe cumulate**
0	20	10	10
1	50	25	35
2	100	50	85
3	15	7.5	92.5
4	10	5	97.5
5	5	2.5	100
	N=200	100%	

Ponderea sau proporția gospodăriilor care au 3 copii sau mai puțin este de 92.5% (10+25+50+7.5=92.5%

D)Pentru a determina abaterea standard vom folosi formula de mai jos:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum(x_i - \bar{x})^2 * f_i}{N-1}}$$

$$s^2 = \frac{(0-1.8)^2*20+(1-1.8)^2*50+(2-1.8)^2*100+(3-1.8)^2*15+(4-1.8)^2*10+(5-1.8)^2*5}{200-1}$$

$$s^2 \ = \frac{222}{199} = 1.11$$

Abaterea standard este radical din varianță:

$s = \sqrt{s^2} = \sqrt{1.11} = 1.05$

E) Datele problemei pe care le cunoaștem sunt următoarele:

n = 200 gospodării

$\bar{x}$ = 1.8

s = 1.05

Media la nivelul populației se încadrează în următorul interval

μ ϵ [$\bar{x}$ – z*e ; $\bar{x}$ + z*e] unde z pentru pragul de 95% este 1.96, iar

$e = \frac{\sigma}{\sqrt{n}} = \frac{1.05}{\sqrt{200}} = 0.0742$

μ ϵ [1.8 – 1.96*0.0742; 1.8 + 1.96*0.0742]

μ ϵ [1.8 – 0.1454; 1.8 + 0.1454]

μ ϵ [1.65; 1.94]

Sunt 95% șanse ca media numărului de copii din gospodărie să ia una dintre valorile din intervalul cuprins între 1.65 și 1.94.

Aplicații de sinteză propuse

Problema 1. Se dau următoarele date, reprezentând distribuția unui eșantion format din 250 gospodării, după numărul persoanelor ce le compun:

Număr de persoane din gospodărie (valorile variabilei)	Număr de gospodării (frecvențe absolute)
1	30
2	40
3	60
4	50
5	30
6	20
7	10
8	5
9	5
Total	250

Determinați indicatorii tendinței centrale.

Determinați ponderea sau proporția gospodăriile care sunt formate din 8 persoane.

Determinați ponderea sau proporția gospodăriile care sunt formate din 6 persoane sau mai puțin.

Determinați valorile quartile (Q1, Q2, Q3).

Problema 2. Se dau următoarele valori, reprezentând notele obținute la statistică de către un grup de studenți ai Facultății de Științe Socio-Umane:

Notele obținute (valorile variabilei)	**Număr de studenți (frecvența absolută)**
4	15
5	20
6	30
7	40
8	7
9	3
10	5
Total	120

Calculați indicatorii tendinței centrale.
Determinați ponderea (proporția) studenților care au luat nota 9.
Determinați ponderea studenților care au luat nota 7 sau mai puțin.
Conform unor date existente la secretariatul facultății media notelor la statistica a studenților este 6.75. Să se testeze semnificația diferenței dintre media notelor obținută la nivel de eșantion (a celor 120 studenți) și media notelor existente la secretariatul facultății.

Problema 3. Din totalul studenților dintr-un centru universitar, s-a format un eșantion aleator de 1600 de studenți. Caracteristica urmărită a fost intenția de a consuma droguri. Rezultatele au fost următoarele:

Intenția de a consuma droguri	Număr de studenți
DA	110
NU	1490
Total	1600

Estimați prin interval de încredere care este proporția celor care au intenția de a consuma droguri la nivelul populației statistice cu o probabilitate de 95%.

Conform unui sondaj efectuat după aplicarea unor programe antidrog în centrul universitar respectiv, proporția celor care au intenția de a consuma droguri a fost de 5.5%; sondajul a fost realizat pe un eșantion format din 1444 studenți. Să se testeze dacă este vorba despre o scădere reală a proporției celor care au intenția de a consuma droguri (cele 2 valori provin din două eșantioane diferite).

Resurse bibliografice recomandate

Babbie, E. (2010). *Practica cercetării sociale*, Iaşi: Polirom.

Caragea, N., Alexandru, C. (2018). *Statistică - concepte, tehnici și instrumente softwaRe*, București: Pro Universitaria.

Clocotici, V., Stan, A. (2000). *Statistica aplicată în psihologie*, Iași: Polirom.

Coman, C. (2011). *Statistica aplicată în științele sociale*, București: Institutul European.

Drugaș, M și Roșeanu, G. (2010). *Analiza statistică pas cu pas: sinteze teoretice, exerciții și demonstrații*, Oradea: Editura Universității din Oradea.

Goodwin, C.James, (2008) *Research in Psychology. Methods and design*, Fifth Edition, USA: John Wiley & Sons, Inc.

Jaba, E., Grama, A. (2004). *Analiza statistică cu SPSS sub Windows*, Iaşi: Polirom.

Kent, R. (2015). *Analyzing Quantitative Data. Variable-based and Case-based Approaches to Non-experimental Data*, Washington DC: Sage Publications.

Opariuc-Dan, C. (2009). *Statistica aplicată în științele socio-umane: noțiuni de bază - statistici univariate*, Cluj-Napoca: Editura ASCR & COGNITROM.

Pop, L. (2002) *Statistică*, (duport de curs), Universitatea București.

Reisz, R. (2017). *Carte de statistică - rețete încercate*. București: Tritonic.

Rotariu, T.(coord.), Bădescu, G., Culic, I., Mezei, E., Mureşan, C. (2006) *Metode statistice aplicate în științele sociale*, Ed. a 2-a, Iași: Polirom.

Săveanu, S. (2020) *Prelucrarea și analiza datelor sociale. Utilizarea programului Excel*, Cluj-Napoca: Presa Universitară Clujeană.

Sora, V., Hristache, I., Mihăescu, C. (1996). *Demografie şi statistică socială*, Bucureşti: Editura Economică.

Șandor, S.D. (2013). *Metode și tehnici de cercetare în științele sociale*. București: Tritonic.

Titan, E., Voineagu, V., Ghiță, S., Boboc, C., Tudose, D. (2007). Statistică - Baze teoretice și aplicații, București: Editura Economică.

Trebici, V. (coord.) (1985). Mică enciclopedie de statistică. București: Editura Științifică și enciclopedică.

Wetcher-Hendricks, D. (2011). *Analyzing Quantitative Data: An Introduction for Social Researchers*, New Jersey: Wiley.

Anexa 1. Tabelul z. Distribuția standard

Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767

2	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Anexa 2. Valorile lui t pentru aria aflată la dreapta

g.l. (grade de libertate)	0.1	0.05	0.025	0.01	0.005	0.001
1	3.078	6.314	12.706	31.821	63.657	318.313
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.182	4.541	5.841	10.215
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.782
8	1.397	1.860	2.306	2.896	3.355	4.499
9	1.383	1.833	2.262	2.821	3.250	4.296
10	1.372	1.812	2.228	2.764	3.169	4.143
11	1.363	1.796	2.201	2.718	3.106	4.024
12	1.356	1.782	2.179	2.681	3.055	3.929
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.610
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.787	3.450
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	3.396
30	1.310	1.697	2.042	2.457	2.750	3.385
31	1.309	1.696	2.040	2.453	2.744	3.375
32	1.309	1.694	2.037	2.449	2.738	3.365
33	1.308	1.692	2.035	2.445	2.733	3.356

34	1.307	1.691	2.032	2.441	2.728	3.348
35	1.306	1.690	2.030	2.438	2.724	3.340
36	1.306	1.688	2.028	2.434	2.719	3.333
37	1.305	1.687	2.026	2.431	2.715	3.326
38	1.304	1.686	2.024	2.429	2.712	3.319
39	1.304	1.685	2.023	2.426	2.708	3.313
40	1.303	1.684	2.021	2.423	2.704	3.307
infinit	1.282	1.645	1.960	2.326	2.576	3.090

Anexa 3. Valorile lui χ^2_{critic} pentru aria aflată la dreapta valorilor

g.l. (grade de libertate)	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401

22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169



ISBN: 978-606-37-1279-1